

# Geometric integration for particle accelerators

Étienne Forest

High Energy Accelerator Research Organization (KEK), 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan

Received 12 September 2005, in final form 7 March 2006

Published 24 April 2006

Online at [stacks.iop.org/JPhysA/39/5321](http://stacks.iop.org/JPhysA/39/5321)

## Abstract

This paper is a very personal view of the field of geometric integration in accelerator physics—a field where often work of the highest quality is buried in lost technical notes or even not published; one has only to think of Simon van der Meer Nobel prize work on stochastic cooling—unpublished in any refereed journal. So I reconstructed the relevant history of geometrical integration in accelerator physics as much as I could by talking to collaborators and using my own understanding of the field. The reader should not be too surprised if this account is somewhere between history, science and perhaps even fiction.

PACS numbers: 02.40.Sf, 02.60.Jh, 29.27.–a, 45.20.Jj

*To the sweet memory of Justin, cruelly taken away*

## Contents

1. Introduction	5322
2. The basic problem and the kick code era	5325
2.1. Kick codes	5327
2.2. TEAPOT: first symplectic integrator for the full $K$	5329
2.3. TEAPOT split	5329
3. Symplectification: good, bad and ugly	5331
3.1. The truncated map of matrix codes	5332
3.2. Problems with the truncated map	5334
3.3. Symplectification by generating functions	5336
3.4. Symplectification by jolts	5340
3.5. Symplectification by other methods	5341
3.6. Generalizing jolts: Abell, Dragt and Rangarajan	5343
3.7. Symplectic restoration	5347
3.8. Methods of restoration for purely linear maps	5350

4. Symplectic integration	5352
4.1. Ruth's integration	5353
4.2. Symplectic modellization or Talman's point of view	5355
4.3. Symplectic integration is restoration	5357
4.4. Yoshida and implications	5358
4.5. Biased integrators and correcting schemes	5360
5. Actual code using symplectic methods	5363
5.1. TEAPOT split reviewed	5364
5.2. The Cartesian bend	5366
5.3. Straight magnets	5366
5.4. The electric elements and non-symplectic radiation	5367
5.5. Fringe fields	5370
5.6. Euclidean group	5372
6. Conclusion and acknowledgments	5374
References	5374

## 1. Introduction

Several years ago, during my late twenties, I had the honour of sharing an office with the late Dr Jackson Laslett, a retired but extremely active scientist at Lawrence Berkeley Laboratory. With Edwin McMillan, René DeVogelaere, Keith Symon and others, he was at the centre of the dynamical discussions surrounding the MURA<sup>1</sup> project of the 1950s. I will quote<sup>2</sup> verbatim Dr Francis T Cole, a man who occupied important positions at Fermi National Laboratory:

Laslett's Work on Chaos

We had seen examples of wandering of phase points and extreme dependence on initial conditions (what would now be called *chaos*) in digital computation as early as 1954. There had always been a nagging uncertainty as to whether the phenomena we saw could be just an effect of computation, although mathematicians were delighted to see the results and said they had been predicting it all along on topological grounds. The Runge–Kutta method we usually used for studying FFAG orbits does not conserve phase space exactly and there are round off and truncation errors as well. Symon and Laslett developed an algebraic transformation [1] that conserved phase space in every term, so non-Liouvillian errors were not present. Laslett, with the help of Storm in the programming, developed this into a dynamics program and studied it extensively. He used double-precision arithmetic to study effects of truncation and tested the roundoff errors by mapping forward through a very large number of steps, then mapping the numerical results backward through the time-reversed system. It took many months of effort to make this system work, because there were extremely subtle roundoff effects that were very hard to cure. Finally, Laslett could demonstrate [2] that there was truly a physical basis to chaos beyond computational imprecision. I believe that this was several years before people working in astronomy achieved the same results. Laslett also discovered the self-similar nature of chaotic motion.

<sup>1</sup> MURA stands for the 'Midwestern Universities Research Association' and was created in 1952. Its mission was to create the next large accelerator in the United States.

<sup>2</sup> This is from an unpublished article of Dr Cole called O Camelot! a memoir of the MURA years, 1994. It can be found on the Web or at the American Institute of Physics, Center for History of Physics, Niels Bohr Library. One Physics Ellipse, College Park, MD 20740, USA. Call Number IH 1995-0021.

Note of the author: FFAG's are fixed field alternating gradient rings and are getting popular again. Remarkably they were invented by the MURA people.

These people were very aware of the Hamiltonian nature of the flow in a proton accelerator. I should say that this is not *a priori* obvious for your randomly selected physicist or engineer. Indeed even if one is aware of the symplectic condition of the map generated by the Hamiltonian  $H$  of special relativity, it is not clear to most physicists that the return map from an arbitrary surface of section leads to a canonical symplectic map in some properly selected coordinates. Without going into mathematical details which are obscure to most accelerator physicists, one can prove that the following extended 1-form (the Langragian after all!) is minimized on the trajectory:

$$\delta \int_{\vec{x}_0, t_0}^{\vec{x}_1, t_1} \vec{p} \cdot d\vec{x} - H dt \Big|_{\delta x, \delta p, \delta t \text{ arbitrary}} = 0. \quad (1)$$

The symmetry of equation (1), particularly the fact that we allow all the positions to vary, makes it plausible that a momentum, let us say  $K = -p_1$ , can be used as a Hamiltonian parameterized by  $x_1$  rather than  $t$ . This new (local) Hamiltonian  $K$  would describe the motion between constant  $x_1$  surfaces. A theorem by G Darboux goes further and proclaims that if more arbitrary surfaces are selected, then some local coordinates can be chosen in which the symplectic forms have the canonical appearance. (For example, the symplectic 2-form is given by  $dq \wedge dp$ ; see [3] for more details.)

So credit must be given to these early pioneers: they knew that ignoring the symplectic condition of the Poincaré map could lead to spurious effects. And, of course, I deserve great discredit for not having had the foresight to discuss these matters more diligently with the likes of Dr Laslett: now that they have crossed the river Styx, unlike Ulysses, I cannot for the sake of this paper descend for a day in the kingdom of Hades. I can still read however and quote Dr Laslett who wrote in January 1985 in the Sardinia [4] conference proceedings:

If one wishes to examine solutions of differential equations, adoption of a 'Hamiltonian' or 'canonical' integration would be reassuring. Such an algorithm has been presented, as a 3<sup>rd</sup>-order algorithm, by R Ruth, and it is understood that Dr Ruth has since developed a similar 4<sup>th</sup>-order integration algorithm—at least for equations derivable from a Hamiltonian function of the form  $H = f(\vec{p}) + V(\vec{q}, t)$ .

Dr Laslett went on to describe early work with McMillan where the Hamiltonian nature was of paramount importance. Simultaneously, as yet unknown to Laslett, Dr Filippo Neri had already generalized the ideas of Ruth. But let me return to Laslett and tell everyone what Dr Ruth reported to me concerning Laslett's contribution which motivated Ruth's own work.

In the early 1980s, Laslett<sup>3</sup> was working on a nonlinear focusing channel. It is well known that a periodic sequence of quadrupoles with alternating gradients produces a linearly stable channel—the so-called FODO cell. However it is not known universally by accelerator physicists that one can realize the same thing with nonlinear multipoles. For example, a continuously rotating multipole creates a stable potential which looks cylindrically symmetric in the rotating frame for small amplitudes and for large rotation rates. Dr Laslett wanted to study these kinds of systems numerically. He was an extremely careful man who documented

<sup>3</sup> Besides the mention of kick codes later in this paper, the reader will note a gap in which characters like Laslett appeared in the 1950s and reappeared in the 1980s. Apparently during this hibernation period, the realization that the Hamiltonian nature of the flow was important was still with us. People knew, for example, that a first-order leap frog was 'better' than an ordinary higher order method when applied to a linear problem in particular. But, the early 1980s were really the beginning of a new area. This is my understanding in part following a conversation with Professor Alex Chao of the Stanford Linear Accelerator.

his work with the minutiae of a Swiss watchmaker. He noted during his simulations small separatrix crossings and other unusual phenomena. This prompted him to contact Ron Ruth who was then a scientist at the Lawrence Berkeley Laboratory. Ruth conjectured that these phenomena were artefacts of the integration method which was a plain Runge–Kutta; a conclusion which did not surprise Dr Laslett given the old MURA stories. This motivated Ruth to look into the development of symplectic integrators. It led to his original paper [5] which contained several integrators including a third-order method which turned out to be a splitting method. Ruth went ahead and discovered a fourth-order method as well during a stay at the CERN laboratory in Geneva. Personally, I lacked interest in Ruth's work because it did not apply to the correct Hamiltonian<sup>4</sup> for ideal magnets<sup>5</sup> and, let us face it, out of intellectual laziness. However when Neri told me that Ruth's third order was a splitting method, I immediately went ahead (1985–1986) and showed that Ruth's fourth order was a splitting method as well. Remarkably none of this was published in any respectable journal. In any event, as I will explain later, I became a full convert to symplectic integration at this point because Ruth's work was then obviously applicable to our most complex standard model. (Professor Richard Talman of SSC/Cornell had also something to do with this as we will later see).

So, this review paper will allow me to fill the gaps and give credit where it belongs. In accelerators, Laslett, Ruth, Neri, Talman, Berz and Dragt contributed directly or indirectly to geometric integration methods. Talman wrote, without realizing it perhaps, the first symplectic integrator for the full 'square-root' Hamiltonian used in our field. Neri noted that Ruth's most famous integrators were splitting methods, thus generalizing their range of applicability. Professors Martin Berz and Alex Dragt provided various tools that allow integrators, symplectic or not, to supplant completely (in my view) the so-called matrix codes (Taylor series codes) even though they are themselves authors of matrix codes.

Personally, I did little as the reader will see. My main contribution was to lump together the work of these gentlemen and try to convince people that integrators (symplectic and nonsymplectic) should supplant the matrix code completely.

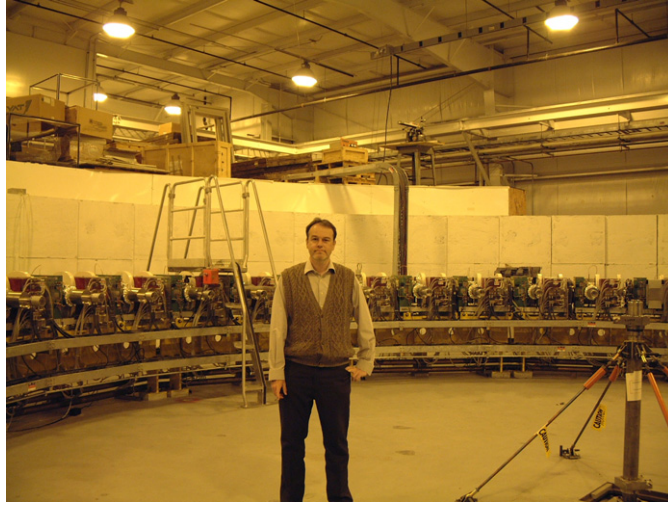
Strangely, the opportunity to write this paper, granted to me by Professors Reinout Quispel and Robert McLachlan despite my poor skills, has a small religious character which results in no small amount from my voluntary exile to Japan and the senseless death of a boy I knew: it has unlocked the pagan wisdom of my pre-Christian ancestors—to provide immortality to good men, not through doubtful beliefs in an afterlife, but by remembering their good deeds and helping others do so as well. Twenty years ago, when these men were alive and willing to transmit their wisdom, I failed miserably to absorb their knowledge and praise their fame. As the Norse poem says

A man must go many places,  
travel widely in the world,  
before he is wise enough to see the workings  
of other men's minds.

So, I am a little wiser and I hope others, despite its glaring shortcomings, will see this paper as an homage to these pioneers of accelerator theory and simulation. These dead men are happy in the Aristotelian sense which is put very succinctly in another Norse poem:

<sup>4</sup> I was still a member of the Taylor map school of thought which emphasizes correct low-order maps all details included. The reader will understand later what this was all about.

<sup>5</sup> The concept of ideal magnets will be mentioned in this paper. While it has at least two different meanings in accelerator physics, here it is more or less a magnet with a constant field in the integration variable  $s$ . Thus these magnets have no fringe fields at the extremities. They are amenable to explicit symplectic integration, i.e., splitting methods.



**Figure 1.** The author appearing lost in the FEL Ring at Duke University.

Cattle die, and kinsmen die,  
And so one dies one's self;  
One thing I know that never dies,  
The fame of a dead man's deeds.

## 2. The basic problem and the kick code era

Since symplectic methods are most relevant to circular storage rings, the reader should think about a storage ring as the prime area of application. Obviously, magnets exist in order to create an ambient electro-magnetic field whose purpose is to guide the particles around a close path akin to a race track. The study of the linear and nonlinear stabilities of this race track, as well as its properties under small parameter changes, is the bread and butter activity of accelerator physicists.

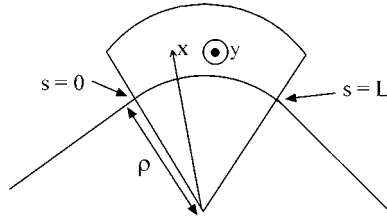
Obviously there exists a global relativistic Hamiltonian that has the form

$$H = \sqrt{m^2 c^4 + c^2 (\vec{p} - q \vec{A})^2} + q \psi. \quad (2)$$

The roles of the various magnets seen in figure 1 are to produce the correct potentials  $\vec{A}$  and  $\psi$ . This description is however very inconvenient since it is usually true that the effect of a magnet is localized in its vicinity. In other words,  $\vec{A}$  and  $\psi$  are nearly zero outside the neighbourhood of a magnet and their values are determined entirely by the local properties of the magnet. Therefore, in analogy with optics, it is useful to view magnets as self-contained objects or lenses which produce a local focusing field quite independently of the world around them. The art of writing a modern design and simulation code resides in the exploitation of this mathematical and physical quasi-object. A first step towards achieving this goal is to use Poincaré surfaces of section roughly perpendicular to the entrance and the exit of a magnet since we intend to study trajectories that traverse this object in one direction<sup>6</sup>.

It is well known to mathematicians that symplectic variables can be found (Darboux) and that we can write a Hamiltonian corresponding to these surfaces of section. Actually we

<sup>6</sup> This is to be contrasted with a cyclotron for example where usually the entire device produces one big electro-magnetic field like one single magnet.



**Figure 2.** Geometry of a sector bend.

can write a Hamiltonian for each individual magnet and we can patch the resulting maps like little bricks. More precisely, the time-like variable is often some length along the direction of propagation. Two transverse coordinates and their associated momenta are supplemented by the energy  $p_t = H$  and the time  $t$ . These form a new canonical pair and thus their Poisson bracket is equal to 1, i.e.,  $[-t, p_t] = 1$ . We remind the reader that in the usual canonical coordinates, the Poisson bracket of two functions  $f$  and  $g$  is given by

$$[f, g] = \sum_{i=1}^3 \frac{\partial f}{\partial x_i} \frac{\partial g}{\partial p_i} - \frac{\partial g}{\partial x_i} \frac{\partial f}{\partial p_i}. \quad (3)$$

For the sake of this paper, I will write a typical Hamiltonian  $K$  in these variables. It describes the motion relative to an arc of circle of radius  $\rho$  in the horizontal plane of a magnet ( $y = 0$ ). When  $\rho^{-1} \neq 0$ , it represents a magnet whose fundamental purpose is to bend the beam. When  $\rho^{-1} = 0$  the magnet is fundamentally a focusing agent.

$$K = -\left(1 + \frac{x}{\rho}\right) \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2} + \frac{x}{\rho} + \frac{x^2}{2\rho^2} + V(x, y; \rho^{-1}), \quad (4)$$

where

$$\vec{z} = (x, p_x, y, p_y, -t, p_t) \quad \text{or} \quad \vec{z} = (x, p_x, y, p_y, p_t, t)$$

is a canonical set of variables<sup>7</sup>. The geometry of the magnet described by equation (4) is illustrated in figure 2.

For the sake of completeness, I should point out that in  $K$  the momenta are scaled by  $p_0$ —the momentum of an ‘ideal’ particle travelling on the circle of radius  $\rho$ . In addition, to keep the algebra simple here,  $p_t$  is actually the total momentum scaled by  $p_0$  and  $t$  is really the path length  $\beta T$ . This implies that this Hamiltonian describes an ultra-relativistic<sup>8</sup> particle. Deviations from the ideal sector bend are contained in the potential  $V(x, y)$ . It should also be added that this Hamiltonian can only represent the body field of an ideal magnet. However, it contains 95% of the physical effects. We will discuss separately the issue of fringe fields and of more realistic fields. See section 3.7.1 for some ideas on fields described numerically and section 5.5 for a discussion of symplectic fringe field models based on impulse calculations.

There are several ways to describe an accelerator on the computer where the integration will be performed. For the purpose of this paper, it suffices to visualize the accelerator as an ordered sequence of mappings representing distinct magnets, including drifts, i.e., regions of free propagation. The surfaces of section are assumed to be smoothly joined. In section 5.6, I

<sup>7</sup> Some computer codes use a positive energy variable which results in a negative time. This is actually very useful in accelerators since the time variable becomes almost a position variable along the beam. Personally, I chose to make  $p_t$  the fifth variables for reasons having to do with automatic differentiation. Dragt uses a positive time and a negative energy. Berz uses yet another definition in his code which is convenient at low energies.

<sup>8</sup> Actually this Hamiltonian describes particles correctly as a function of the total momentum as far as the transverse phase plane is concerned. However, the variable  $t$  is the total path length rather than the time  $cT$ ; this is not satisfactory if the ring contains electric elements capable of changing the energy unless we deal with ultra-relativistic particles.

will briefly discuss the mathematical and programming structures which permit us to envisage more general descriptions for a beam line: although it is a topic disconnected from the main focus of this paper, it is, like Darboux's coordinates, an important element of any attempt to make magnets independent objects dynamically speaking.

With this caveat in mind, a magnet of curvilinear length  $L$  produces a map which is the result of integrating equation (2). Denoting this map  $\xi_i$  for the  $i$ th magnet, the result for a ring or a piece of a ring containing  $N$  magnets (drift spaces included as magnets) is just the concatenation of these  $N$  maps:

$$\xi_{1 \rightarrow N} = \xi_N \circ \cdots \xi_k \circ \cdots \xi_1. \quad (5)$$

How to evaluate this map approximately and symplectically is at the heart of Geometric Methods in accelerators. We will now review the main numerical methods used during what I called the 'kick code era': the original kick codes and the ground breaking code TEAPOT of Talman.

### 2.1. Kick codes

It is obvious that if one can solve exactly the Hamiltonian under consideration, then we have trivially a symplectic integrator. Accelerator physicists in the 1960s and 1970s exploited the fact that the angles with respect to the design orbit are small in large storage rings. Looking at equation (2), we can expand  $K$  in powers of  $p_x$  and  $p_y$ :

$$K_e = -p_t + \frac{p_x^2 + p_y^2}{2(1 + p_t)} - \frac{x p_t}{\rho} + \frac{x^2}{2\rho^2} + V_2 + V_{\geq 3}. \quad (6)$$

In equation (6),  $V_2$  represent the quadratic part of  $V$ . The terms neglected in the momenta are assumed to be smaller than  $V_{\geq 3}$ . It turns out that one must introduce 'sextupole' magnets for reasons that have nothing to do with single particle dynamics. These magnets, with a vector potential proportional to  $x^3 - 3xy^2$ , are set so as to eliminate the linear dependence of the eigenvalues of the linear map on the energy  $p_t$ : the so-called chromaticities. The magnitudes of the resulting sextupoles are such that they become the dominant nonlinear effect. Therefore the expansion in equation (6) was justified for a large class of machines.

In the minds of accelerator physicists, two classes of magnets thus appeared: linear magnets and nonlinear ones. The linear magnets were the bends and the quadrupoles. A pure sector bend would then be governed by the Hamiltonian:

$$K_b = -p_t + \frac{p_x^2 + p_y^2}{2(1 + p_t)} - \frac{x p_t}{\rho} + \frac{x^2}{2\rho^2}. \quad (7)$$

A pure quadrupole would in turn be represented by the Hamiltonian:

$$K_q = -p_t + \frac{p_x^2 + p_y^2}{2(1 + p_t)} + k_Q \frac{x^2 - y^2}{2}. \quad (8)$$

Of course a combined function bend—a bend with quadrupole field—could be represented by

$$K_c = -p_t + \frac{p_x^2 + p_y^2}{2(1 + p_t)} - \frac{x p_t}{\rho} + \frac{x^2}{2\rho^2} + V_2. \quad (9)$$

Now a few remarks are in order:

1. The Hamiltonians  $K_b$ ,  $K_q$  and  $K_c$  are all quadratic and therefore are all *exactly* solvable in terms of matrices as far as the transverse dynamics is concerned.
2. The solution is obviously a linear polynomial map. Therefore these objects became known in accelerator physics as the 'linear magnets'. To this day there are practitioners in our field

who are stunned to learn that these magnets are not truly linear! The torpor has subsided in recent years with the appearance of systems where the small angle approximation is invalid: small colliders and fixed field alternating gradient rings (FFAG). Remarkably the FFAG was conceived by the MURA people in the 1950s.

3. Thus, within the small angle approximation, an ideal ring is a linear object.
4. Nonlinearity enters as one attempt to correct the chromaticities with sextupoles.
5. Sextupoles are thus purely nonlinear in a perfect machine and they can be represented most of the time by a single impulse:

$$K_s = -p_t + \frac{p_x^2 + p_y^2}{2(1 + p_t)} + Lk_s \delta(s - s_0) \frac{x^3 - 3xy^2}{3}. \quad (10)$$

Here the variable  $s$  is a length along the direction of the magnet ( $K_s = -p_s$ ). The sextupole is centred around  $s = s_0$ . The exact solution for a sextupole using the Hamiltonian of equation (10) is of the form,

$$\xi = d \circ k \circ d, \quad (11)$$

where the map  $d$  is a free propagation of length  $L/2$  and  $k$  is an impulse or the so-called kick:

$$k \rightarrow p_x^{\text{final}} = p_x - Lk_s(x^2 - y^2) \quad \text{and} \quad p_y^{\text{final}} = p_y + Lk_s 2xy. \quad (12)$$

Most readers will recognize that this is just one step of the leap frog symplectic integrator often called the drift-kick method in accelerator circles.

6. In addition, the reader will note that the change of the time  $t$  through the device is given by a quadrature over the solutions for the transverse motion:

$$t^{\text{final}} = \int_0^L \left\{ 1 + \frac{p_x^2(s) + p_y^2(s)}{2(1 + p_t)^2} + \frac{x(s)}{\rho} \right\} ds. \quad (13)$$

The fact that nonlinear elements such as sextupoles were represented by impulses as in equation (12) gave rise to the appellation ‘kick codes’.

Perhaps one of the first kick codes was the code SYNCH developed by Al Garren for which he received a prize a few years ago. The prize was for its usage of what accelerator physicists call the Courant–Snyder [6] theory. In fact it is a Floquet description taking advantage of the periodicity of the system. SYNCH was the first design code using this Floquet theory by tracking the invariant ellipses around the ring using the matrices which naturally emerge from the linear elements. SYNCH was not really a simulation code and contained, at least in the 1980s, various ad hoc formulae whose purposes were to provide corrections to the linear matrices: restoring the expanded square-root and adding some fringe field effects. Ring design was truly SYNCH’s initial function. Programming and development involved many people besides Garren including the famous Ernest Courant co-inventor of strong focusing.

Later, several kick codes were written for the study of stability. One such code, called PATRICIA [7, 8], written by Weidemann of SLAC, used the transverse formalism for tracking but failed to provide a coherent calculation of the time  $t$ . It is often the case in accelerators that people forget that time is a canonical variable and thus revert to more ad hoc formulae for its computation.

Amongst the first persons to include a correct description of the time computation were Drs G Ripken, F Schmidt and D P Barber. They extended a DESY code called RACETRACK [9] of A Wrulich which was functionally a transverse code like PATRICIA. To this code they added, amongst other things, the quadrature for the computation of the time. The resulting modified code was rebaptized SIXTRACK [10, 11], obviously the ‘SIX’ referring to its proper

description of the full six-dimensional phase space. Dr Ripken, who left us recently, was an expert in Hamiltonian methods.

The list of codes using this kick code formalism is endless. It is used in codes with primary functions sitting outside single particle dynamics, for example codes describing collective effects. Simply put, accelerator physicists intuitively trusted this model because it was an ‘exact solution’ to an approximate and sufficiently realistic model. Of course, by the time Neri and I examined Ruth’s seminal work, we realized that kick codes were simply symplectic integrators using a particular quadratic split with a single step of integration!

Nevertheless, late in the 1980s, there was still a ‘practical’ distinction made between advocates of ‘kick codes’ and ‘symplectic integration’. Of course, kick codes are symplectic integrators.

## 2.2. TEAPOT: first symplectic integrator for the full $K$

TEAPOT [12] is the first code to integrate the full Hamiltonian  $K$  in a symplectic manner. The original paper on TEAPOT is quite confusing about the entire issue of ‘symplecticity’ except for the fact that having solved a Hamiltonian problem exactly, the result ought to be symplectic, whatever the hell that means! Notwithstanding the confusion, they were absolutely correct: their integrator was a second-order symplectic integrator adequate for the full Hamiltonian  $K$ .

As it often happens in this incestuous community, I was contacted by one of the reviewers and by the main author, Professor Richard Talman, who happened to be working at the defunct SSC Central Design group where the Lawrence Berkeley Laboratory, my employer, assigned me as a new and incompetent accelerator physicist. The reviewer’s blunt question was simply: ‘How is this different from kick codes?’ In other words, it looked pretty much to him like  $K_e$  of equation (6). This impression was compounded by claims in the TEAPOT paper concerning large machines: a thin lens integrator<sup>9</sup> with relatively few integration steps (one per magnet for 99% of the SSC!) should be adequate in a large beast like the SSC. In fact, they were pulling the small angle arguments once more to justify the small number of steps—the same argument which justified the expansion used to produce  $K_e$ .

When an incensed Talman showed me the comments of the reviewer, I contributed an appendix to [12] where I showed that TEAPOT could produce, if the number of steps per magnet were increased, the correct results for a strong bending magnet. In other words, the additional work that was put into TEAPOT to integrate the full Hamiltonian  $K$ , was really relevant to small storage rings and not to grotesque rings like the SSC or the present LHC at CERN. In fact, most standard rings can be simulated with the expanded  $K_e$  quite adequately.

Nevertheless, TEAPOT is indeed the first symplectic integrator for the full Hamiltonian  $K$  used in accelerator physics. I do not want here to go over the original derivation which uses noncanonical variables. Instead I will describe TEAPOT with modern terminology.

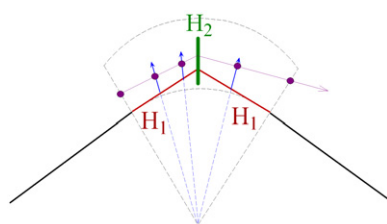
## 2.3. TEAPOT split

The Hamiltonian is divided as follows:

$$K = \underbrace{-\left(1 + \frac{x}{\rho}\right) \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}}_{\text{drift in polar coordinates}} \underbrace{+ \frac{x}{\rho} + \frac{x^2}{2\rho^2} + V(x, y)}_{\text{multipole kick}}. \quad (14)$$

TEAPOT<sub>split</sub>

<sup>9</sup> One more word for a single kick or a single step of a leap frog method!



**Figure 3.** Geometry of TEAPOT integration step.

This idea behind splitting methods is that each pieces, namely  $H_1$  and  $H_2$  are exactly solvable. For example the solution for  $H_1$ , a drift in polar coordinates, is given by equation (127). But here I will emphasize the intuitive approach initially adopted by Talman.

One can see from figure 3 that it is possible to view a single step of symplectic integration as three ‘physical’ processes done consecutively. First, there is a drift, i.e., a free propagation. This free propagation, in the Hamiltonian context, must be done in polar coordinates. Thus, as shown on figure 3, the frames rotate as the particle propagates under the effect of  $H_1$ . In the middle of the step, it encounters an impulse which deflects the trajectory. The particle then propagates once more in the polar drift  $H_1$ .

Originally, this integrator was derived using a geometric picture and a solution of Lorentz’s equation for the impulse. The drifts were done using geometry: the intersection with the thin lens  $H_2$  was computed. At that point a rotation was applied to put the ray (the directions  $x'$  and  $y'$ ) in coordinates parallel to the thin lens. On the other side of the thin lens, an identical rotation was performed, followed by a regular drift in Cartesian coordinates. One realizes that these operations, viewed by Talman separately, could never be ‘symplectic’ in the standard sense. But, of course, if done correctly, they would lead to a second-order symplectic integrator for  $K$  in a proper set of coordinates.

Moreover, as part of this intuitively derived scheme, Talman considered the straight lines to be the true design trajectory of the particle. Thus the machine in TEAPOT was truly made into a polygon. This is an extreme view of integration found in accelerator physics: we create models that become reality in the highest sense of the word. Therefore, modellization is what we are doing rather than *bona fide* integration. It should be pointed out that TEAPOT used typically one step per magnet, as displayed in figure 3, and that the various parameters of the potential were adjusted to achieve the desired properties including a closing polygon! This approach and usage of symplectic methods is typical. Here it reached extremes which are not empirically needed.

Nevertheless Talman did create the first symplectic integrator for the Hamiltonian  $K$ . Both the ancient kick codes of section 2.1 and TEAPOT were not really connected to Ruth’s algorithm since Ruth had derived his formulae for a Hamiltonian of the form  $T(p) + V(q, t)$ . Therefore, these achievements were completely independent. The necessity of a symplectic scheme, a realization whose popularization, dare I say vulgarization, is more due to Dragt than Ruth, and lethal bugs found in codes used at the SSC Central Design Group were the driving force behind what Talman called ‘an exact scheme’: an exact solution to a simple non-expanded model. I strongly objected at that time on the necessity of an integrator for the full Hamiltonian  $K$  on the grounds that it was not necessary for the SSC. Perhaps Talman would agree now with me on this, but I am now happy that he did not in those days and that I did not have the clout to stop him. He contributed a lot to the spreading of symplectic integration by teaching us that the full Hamiltonian of equation (4) can be split into solvable parts.

This brings us to Ron Ruth and Alex Dragt. I would like to describe their respective contribution in some detail since it has an impact on the entire usage of symplectic integration in the field of storage rings.

### 3. Symplectification: good, bad and ugly

Historically, in the early 1980s, I personally got involved with a procedure called ‘symplectification’. Certain approximate maps, generally Taylor series maps, were computed for a piece of the machine or even an entire machine. The maps originated from so-called Taylor series matrix codes. In those days, these codes would store for each magnet a formula for the Taylor expansion of a mapping to second or third order around some ideal orbit. Maps were then concatenated to form a global map for a large section of the ring. These maps were then used carelessly in tracking studies and produced spectacularly nonsymplectic results. Of course this is not surprising since they represented the truncated power series of a true symplectic map.

Matrix codes had been around for years. The most famous matrix code was called TRANSPORT [13] and was the brain child of Karl Brown of SLAC. They were used mainly to design single pass systems such as spectrometers. In these systems, the entire apparatus of geometric integration is rather useless. These systems are like cameras or electron microscopes: no one iterates the map.

The procedure of symplectification consists in completing the power series by adding extra terms in a rather ad hoc way. Therefore the effects of the additional terms could not be fully trusted beyond a certain radius<sup>10</sup> in phase space. Of course, whether or not one can attach some credence to these symplectified maps, depends on the studies themselves. Unfortunately these symplectified maps were used to study the dynamic aperture of machines: the very area where our knowledge of the Taylor map is completely nil. It became clear to some, in particular Ron Ruth and Dick Talman (see footnote 24), that this type of uncontrolled usage of the ‘symplectification’ algorithms was dangerous and oversold by their proponents, including myself. For this reason I will define two limiting types of symplectification: *map fabrication* and *map restoration*. What I just briefly mentioned is Taylor fabrication.

So, in counterpoint literally to the symplectification effort, Ruth went ahead in searching for an improvement to the second-order methods which were in vogue in accelerator physics, namely the kick codes. Ruth later said that, when one agrees to a certain model for the transverse Hamiltonian  $K$ , uncontrolled approximations in the transverse directions are very dangerous while longitudinal ones, as long as they preserve the symplectic nature of the flow, are far less dangerous. The longitudinal variable is our integration variable  $s$ , therefore better symplectic approximations in the step size  $ds$  are what we call now symplectic integration schemes. If we examine the Taylor series of a map produced through symplectic integration, the various orders converge towards the exact result for the chosen Hamiltonian  $K$  more or less in unison: for this reason symplectic integration techniques fall within the domain of map restoration as illustrated in figure 14. However, we will see that symplectification of the Taylor series can also be an exercise in restoration rather than a pure act of fabrication.

<sup>10</sup> The radius of convergence of the Taylor series is limited in the energy variable  $p_t$  to values of the order of one (100%). This is unfortunate for proponents of Taylor codes particularly if the beam is accelerated as it is done in FFAG. In the transverse plane the Taylor series seems acceptable to radii equal or exceeding the dynamic aperture. Indeed if one believes, as most accelerator physicists do, that the so-called kick codes using the expanded Hamiltonian are appropriate in ‘large’ machines, then there are no fundamental problems with the Taylor maps since the kick codes are Taylors series! The integration of equation (6) with a leap frog method will give a polynomial map since  $V_{\geq 3}$  is usually limited to a polynomial. Paradoxically this is what permits codes using very high-order Taylor maps, such as COSY INFINITY [14] of Martin Berz to have some hope of working.

In this paper, I will review both symplectic integration schemes and symplectification schemes using simple examples, because it is important to distinguish symplectification (fabrication and restoration) from symplectic integration (normally a pure restoration method).

### 3.1. The truncated map of matrix codes

To understand the context which gave rise to symplectification we must first define the so-called matrix code. These codes existed in parallel with the kick codes during the 1970s. As I said the most famous<sup>11</sup> one amongst accelerator physicists is probably TRANSPORT, a code [13] written by Dr Karl Brown of SLAC. The idea of the standard matrix codes is to expand around the so-called design trajectory. In the case of straight elements, it is a straight line right down the middle of the magnet. For bends, it is typically a circle of radius  $\rho$ .

Generally speaking, the code will represent each magnet using a Taylor series expansion around the design trajectory:

$$\begin{aligned} z_a^{\text{final}} &= \sum_b T_{ab}^1 z_b^0 + \sum_{b \leq c} T_{abc}^2 z_b^0 z_c^0 + \sum_{b \leq c \leq d} T_{abcd}^3 z_b^0 z_c^0 z_d^0 + \dots \\ &= \sum A_{j_1 \dots j_6}^a z_1^{j_1} z_2^{j_2} z_3^{j_3} z_4^{j_4} z_5^{j_5} z_6^{j_6}. \end{aligned} \quad (15)$$

The ‘matrices’  $T^i$  or equivalently  $A_j^a$  contain the Taylor coefficients for each individual elements. Originally, in the 1970s, analytic formulae were derived. TRANSPORT was second order (i.e.  $T^2$ ), MARYLIE, a code storing the Taylor series using a Lie representation, had formulae through  $T^3$ . The record for analytically derived Taylor code belongs to COSY 5.0 of Professor Martin Berz [15]. Nowadays, using the techniques of automatic differentiation, we can compute routinely these matrices to very high order. This is what the code COSY INFINITY [14] of Berz does systematically.

Using our Hamiltonian  $K$ , we can look at an example of a ‘matrix code’. For the sake of the example, let us assume that we have straight elements ( $\rho = 0$ ) and that we are living in one dimension. Consider a generic one-dimensional quadrupole–sextupole of strength given by  $k_Q$  and  $k_S$  respectively and described approximately by the following  $K$ :

$$K = -\sqrt{1 - p_x^2} + \frac{k_Q}{2} x^2 + \frac{k_S}{3} x^3. \quad (16)$$

A matrix code which computes third-order Taylor maps will expand this Hamiltonian to fourth order around the central axis  $(x, p_x) = (0, 0)$ :

$$K_4 = \frac{p_x^2}{2} + \frac{p_x^4}{8} + \frac{k_Q}{2} x^2 + \frac{k_S}{3} x^3. \quad (17)$$

The next step consists in computing the exact Taylor series for the map produced by  $K_4$  to third order in the vector  $z = (x, p)$ . Let us do this for the semi-realistic lattice of figure 4.

The parameters of each individual elements are given in figure 4. The total map for this cell<sup>12</sup> is

$$\xi = D \circ S_d \circ Q_d \circ D \circ S_f \circ Q_f. \quad (18)$$

In equation (18),  $Q_f$  and  $Q_d$  are the quadrupoles of figure 4 while  $S_f$  and  $S_d$  are the sextupoles. The map  $D$  is a simple drift of length  $L = 2.0$ .

<sup>11</sup> The theory, described in [13], is so famous in accelerator circles that it is often quoted using only its report number: SLAC-75!

<sup>12</sup> The term ‘Fodo cell’ refers to the focusing–drift–defocusing–drift nature of the cell. The author used to believe that this structure was named after Mr Fodo—a reassuring thought to the readers taking this paper seriously!

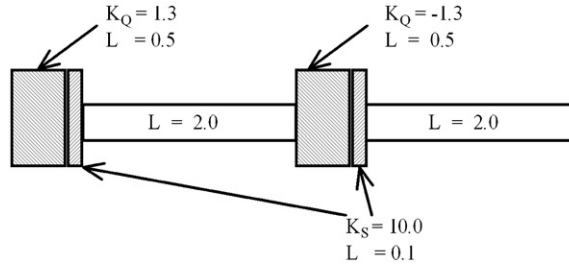


Figure 4. Sketch of a FODO cell.

Let us expand this map as a Taylor series through second order. The result is to machine precision:

$A(1, i, j)$	$i \ j$	$A(2, i, j)$	$i \ j$
-3.006 853 442 731 210	1 0	-1.026 971 302 129 112	1 0
8.350 287 455 798 908	0 1	2.519 412 976 361 477	0 1
-7.146 474 264 408 377	2 0	-2.507 623 185 315 878	2 0
4.824 310 130 239 798	1 1	3.599 555 856 289 304	1 1
-23.013 774 435 006 73	0 2	-10.837 220 637 324 01	0 2

Here, as in equation (15), the matrix  $A(1, i, j)$  gives us the Taylor expansion of the final position  $\xi_1$  in terms of the initial coordinates.  $A(2, i, j)$  gives us the momentum  $\xi_2$ .

For our discussion, it is convenient to transform the map by a linear canonical similarity transformation  $a$  so that the linear map becomes a pure rotation. The coefficients<sup>13</sup> of the map  $a$  are given by

$A(1, i, j)$	$i \ j$	$A(2, i, j)$	$i \ j$
2.934 265 570 260 178	1 0	0.970 956 591 103 2019	1 0
		0.340 800 781 679 5298	0 1

The final map  $\zeta$  is obtained by similarity transformation:

$$\zeta = a^{-1} \circ \xi \circ a. \quad (19)$$

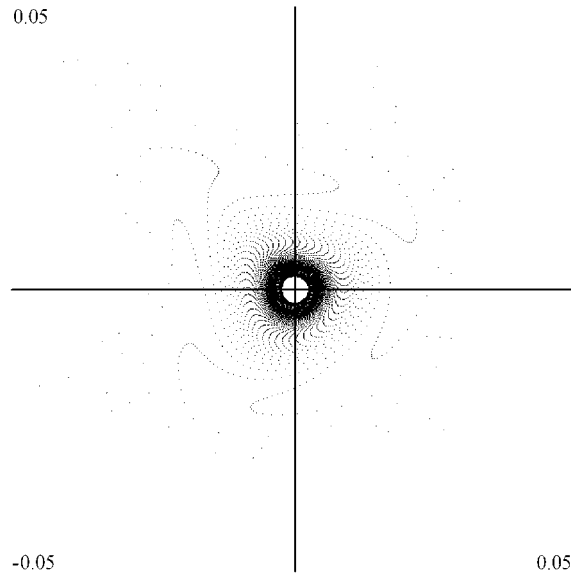
The resulting coefficients for the map  $\zeta$  are

$A(1, i, j)$	$i \ j$	$A(2, i, j)$	$i \ j$
-0.243 720 233 184 8667	1 0	-0.969 845 579 428 1455	1 0
0.969 845 579 428 1458	0 1	-0.243 720 233 184 8673	0 1
-23.679 604 364 051 93	2 0	4.224 988 197 896 419	2 0
-3.546 486 443 777 293	1 1	-0.378 791 115 782 4233	1 1
-0.910 939 635 278 0329	0 2	-1.098 025 710 819 811	0 2

(20)

This map can be used to track the lattice described in figure 4. The phase-space portrait, obtained by repeatedly applying this map, is displayed in figure 5. This figure shows a single trajectory starting at the centre and diverging towards the outside. This is not a physical result but a consequence of the truncated nature of the map  $\zeta$ .

<sup>13</sup> The map  $a$  is the so-called Courant–Snyder transformation. It is defined uniquely up to a rotation in normalized phase space. Accelerator physicists usually select the rotation so that  $A(1, 0, 1)$ , the dependence of position on normalized momentum, is zero. This choice is convenient in analytic calculations if nonlinear terms depend only on position.



**Figure 5.** Phase space of a second-order truncated map.

### 3.2. Problems with the truncated map

There are two reasons why a truncated map fails to be symplectic. The first obvious reason is truncation and the second is related to inexact Taylor coefficients. Traditionally the Taylor coefficients in a code such as TRANSPORT or MARYLIE were obtained analytically and therefore the coefficients themselves were absolutely correct notwithstanding various programming and mathematical errors in the first versions of these codes. In the example of figure 5, we can check that the coefficients in equation (20) obey the symplectic condition in a way consistent with the truncation order:

$$[\zeta_1, \zeta_2] = 1.0 + 0.8910^{-14}z_1 + 0.2710^{-14}z_2 + 47.9z_1^2 + 119.4z_1z_2 + 7.1z_2^2. \quad (21)$$

I left the numbers as they appeared on my computer. For a truly symplectic map the answer should have been one. We can see that the tiny coefficients in the linear part of this polynomial are actually due to floating point roundoff errors. The quadratic part, far from being zero, is a direct consequence of the truncation to second order. This violation of the symplectic condition is directly responsible for the unpleasant antidamping in figure 5.

In the early 1980s, as many practitioners of the voodoo art of accelerator design tried to do simulations using codes such as TRANSPORT, perplexing phase plots started to appear, which for good reasons, had not been seen in ‘kick codes’. People such as Karl Brown could not be blamed for this: they never intended that their codes be used for ring simulation.

However, at that time, Alex Dragt and his students were working on the code MARYLIE and various Lie algebraic tools with direct relevance to circular accelerators. MARYLIE, being an expansion to third order, was also susceptible to nonsymplectic behaviour. In MARYLIE, the Taylor series representation is parameterized by Lie operators. The expansion of equation (15), which is a third-order expansion, corresponds to MARYLIE 3.0, a code available

for download on the Internet. We can rewrite this expansion by factoring the linear part out. For example, the map  $\zeta$  is rewritten as follows:

$$\zeta = N \circ L. \quad (22)$$

The map  $L$  is the linear part:

$$\begin{array}{cc} A(1, i, j) & i \ j \\ -0.243 \ 720 \ 233 \ 184 \ 8667 & 10 \\ 0.969 \ 845 \ 579 \ 428 \ 1458 & 01 \end{array} \quad \begin{array}{cc} A(2, i, j) & i \ j \\ -0.969 \ 845 \ 579 \ 428 \ 1455 & 10 \\ -0.243 \ 720 \ 233 \ 184 \ 8673 & 01 \end{array} \quad (23)$$

The map  $N$ , the nonlinear part, has the Taylor coefficients:

$$\begin{array}{cc} A(1, i, j) & i \ j \\ 1.000 \ 000 \ 000 \ 000 \ 000 & 1 \ 0 \\ -1.425 \ 101 \ 606 \ 781 \ 972 & 2 \ 0 \\ -7.638 \ 536 \ 649 \ 513 \ 914 & 1 \ 1 \\ -23.165 \ 442 \ 392 \ 547 \ 96 & 0 \ 2 \end{array} \quad \begin{array}{cc} A(2, i, j) & i \ j \\ 1.000 \ 000 \ 000 \ 000 \ 000 & 0 \ 1 \\ -0.692 \ 305 \ 837 \ 680 \ 3470 & 2 \ 0 \\ 2.850 \ 203 \ 213 \ 563 \ 945 & 1 \ 1 \\ 3.819 \ 268 \ 324 \ 756 \ 953 & 0 \ 2 \end{array} \quad (24)$$

The map  $L$  being linear is exactly symplectic and does not present any problem. The idea in MARYLIE was to represent the nonlinear part  $N$  using a Lie operator. Consider a cubic polynomial  $f_3$  in the variable  $z = (z_1, z_2)$ , then a second-order map without any linear terms, can be represented as

$$\begin{aligned} N &= \exp(:f_3:)I = I + [f_3, I] + \dots, \\ :f_3: g &= \underbrace{[f_3, g]}_{\text{Poisson bracket}}, \\ \text{where } I &= \text{identity map.} \end{aligned} \quad (25)$$

The map  $N$  can be viewed as the map generated<sup>14</sup> by the Hamiltonian ‘ $-f_3$ ’ for a pseudo-time of  $\tau = 1$ .

The quadratic part of the Taylor coefficients of  $N$  can be directly linked to  $f_3$ :

$$\begin{aligned} \text{if } f_3(z_1, z_2) &= F_{30}z_1^3 + F_{21}z_1^2z_2 + F_{12}z_1z_2^2 + F_{03}z_2^3 \\ [f_3, I](z_1, z_2) &= (-F_{21}z_1^2 - 2F_{12}z_1z_2 - 3F_{03}z_2^2, 3F_{30}z_1^2 + 2F_{21}z_1z_2 + F_{12}z_2^2). \end{aligned} \quad (26)$$

One notes that equation (26) implies that certain relations must exist between the coefficients of the Taylor series:

$$A_{120} + \frac{1}{2}A_{211} = 0 \quad \text{and} \quad \frac{1}{2}A_{111} + A_{202} = 0. \quad (27)$$

These relations are satisfied to machine precision in equation (24). This is the case for matrix codes and Taylor maps extracted from symplectic integrators: kick codes, TEAPOT and modern integrators.

The idea of Alex Dragt and his student David Douglas was to parameterize Taylor maps in terms of Lie polynomials. For example, a second-order code necessitates a cubic polynomial  $f_3$ . The code MARYLIE 3.0, a third-order code, requires  $f_3$  and  $f_4$ :

$$\begin{aligned} N &= \exp(:f_3:) \exp(:f_4:)I \\ &= I + [f_3, I] + [f_4, I] + \frac{1}{2}[f_3, [f_3, I]] + \dots \end{aligned} \quad (28)$$

<sup>14</sup> The operator  $: -f :$  is often denoted by  $L_f$  in the literature. Here I stick to Dragt’s notation for the sake of those accelerator physicists who have not yet given up reading this paper at this junction.

Although this is ‘officially’ symplectic, in reality it is only symplectic in a simulation code if one can evaluate an infinite number of Poisson brackets. Unable to do this, Dragt and his group started a search for methods that could be used to evaluate a map in a fully symplectic manner.

### 3.3. Symplectification by generating functions

There are several ways to turn an approximate map into a true symplectic map. We first describe a method developed by Dragt, Douglas and Neri for maps parameterized using the Dragt–Finn [16] Lie factorization. Later it was applied to low-order Taylor maps by Douglas, Forest and Servranckx [17]. Finally Berz, for Taylor maps, programmed tools to compute generating functions for Taylor maps to arbitrary order. Berz provided a partial Taylor map inverter which allows one to compute a mixed variable map which must be the gradient of a generating function. This inverter is part of Berz’s automatic differentiation package and his code COSY INFINITY [14]. Berz and Erdelyi, in [18], have also advocated a Poincaré type of a generating function which is computable easily from the Taylor map. (See below the discussion surrounding equation (111).)

**3.3.1. Generic formulae for low orders.** We will describe here the original method of Dragt, Douglas and Neri because it is not restricted to Taylor maps. It applies to generic maps factored and ordered using a smallness parameter  $\varepsilon$ . Consider a map  $N$  produced by the following Lie representation:

$$\begin{aligned} N &= \exp(\varepsilon f_1) \exp(\varepsilon^2 f_2) \cdots \exp(\varepsilon^\nu f_{N_0}) \\ &= I + \varepsilon [f_1, I] + \varepsilon^2 \left\{ [f_2, I] + \frac{1}{2} [f_1, [f_1, I]] \right\} \\ &\quad + \varepsilon^3 \left\{ [f_3, I] + [f_1, [f_2, I]] + \frac{1}{6} [f_1, [f_1, [f_1, I]] \right\} \\ &\quad + \cdots \end{aligned} \quad (29)$$

This factorization [16] is known in accelerator circles as the ‘Dragt–Finn’ factorization. It appears naturally in perturbation theory and was also chosen in the code MARYLIE as the standard way to describe a map—magnet or beam line.

The idea of Dragt and collaborators was to rewrite equation (29) as a mixed variables generating function. Consider  $G$  defined as

$$\begin{aligned} G(q^{\text{in}}, p^{\text{final}}) &= q^{\text{in}} \cdot p^{\text{final}} - \sum_{i=1}^{\nu} \varepsilon^i G_i(q^{\text{in}}, p^{\text{final}}) \\ &\Downarrow \\ q^{\text{final}} &= q^{\text{in}} - \sum_{i=1}^{\nu} \varepsilon^i \frac{\partial G_i}{\partial p^{\text{final}}}(q^{\text{in}}, p^{\text{final}}) \\ p^{\text{final}} &= p^{\text{in}} + \sum_{i=1}^{\nu} \varepsilon^i \frac{\partial G_i}{\partial q^{\text{in}}}(q^{\text{in}}, p^{\text{final}}). \end{aligned} \quad (30)$$

One must find formulae, order by order in  $\varepsilon$ , until the map generated by equation (30) agrees to order  $\varepsilon^\nu$  with the factored Lie map  $N$  of equation (29). Then, as is done with implicit Runge–Kutta, the implicit equation (30) is solved to machine precision on the computer.

For example, we can do the trivial case of  $\nu = 1$  which corresponds to our cell example and figure 5. This requires us to equate equations (29) and (30) and solve to first order in  $\varepsilon$ . We get the trivial result:

$$G_1 = f_1. \quad (31)$$

In the code MARYLIE 3.0, equation (31) was extended to second order by Douglas. The results are

$$G_1 = f_1, \quad G_2 = f_2 - \frac{1}{2} \frac{\partial f_1}{\partial q} \cdot \frac{\partial f_1}{\partial p}. \quad (32)$$

As we can see in figure 7, the inclusion of one additional degree in the Taylor representation and in the generating function improves the overall phase plot.

*3.3.2. High-order formulae: Douglas and Neri's graphs.* Generating functions can be computed easily to arbitrary order when dealing with a Taylor map using partial inverters such as that found in Berz's Taylor series package. However historically formulae in MARYLIE were computed for a generic representation of the map. Initially this calculation was started by Douglas in February 1982 through order  $\nu = 2$  and implemented in MARYLIE in early 1983. This corresponds to cubic terms in the Taylor series of MARYLIE (equation (32)). Douglas later extended these formulae through fourth order ( $\nu = 4$ ) most likely around fall 1983 or early 1984. In the summer of 1983, I used the formulae of equations (31) and (32) in connection with the code TRANSPORT while at Los Alamos National Laboratory<sup>15</sup>.

These painfully derived fourth-order formulae were derived simultaneously (or soon after) by Neri using a home-grown calculus that could be implemented in algebraic manipulators. As seems to be the norm with Neri's work, this work is unpublished as far as I know. I suppose that one can get the formula from Alex Dragt in Maryland since this work of Neri was incorporated in an experimental version of the code MARYLIE called MARYLIE 5.0. In my dusty archives, I found a photocopy of Neri's derivation of  $G_4$  in terms of the  $f_i$ 's. The gruesome result is

$$\begin{aligned} G_4 = & f_4 - \frac{1}{2} d_i f_2 \delta_i f_2 - d_i f_3 \delta_i f_1 + \frac{1}{2} \delta_i f_1 d_i \delta_j f_1 d_j f_2 + \frac{1}{2} d_i \delta_j f_2 \delta_i f_1 \delta_j f_2 + \frac{1}{2} \delta_i \delta_j f_1 d_i f_2 d_j f_1 \\ & - \frac{1}{24} \delta_i \delta_j \delta_k f_1 d_i f_1 d_j f_1 d_k f_1 - \frac{1}{8} \delta_j \delta_k f_1 d_j f_1 d_k (d_i f_1 \delta_i f_1) \\ & - \frac{1}{24} \delta_k f_1 d_k [\delta_i \delta_j f_1 d_i f_1 d_j f_1 + \delta_j f_1 d_j (d_i f_1 \delta_i f_1)]. \end{aligned} \quad (33)$$

In equation (33), besides the fact that repeated indices are summed over, I adopted the following short hand notation:

$$d_i = \frac{\partial}{\partial q_i}, \quad \delta_i = \frac{\partial}{\partial p_i}. \quad (34)$$

In addition, high-order derivatives are expressed using products of the operators  $d_i$  and  $\delta_i$ . For example, we have

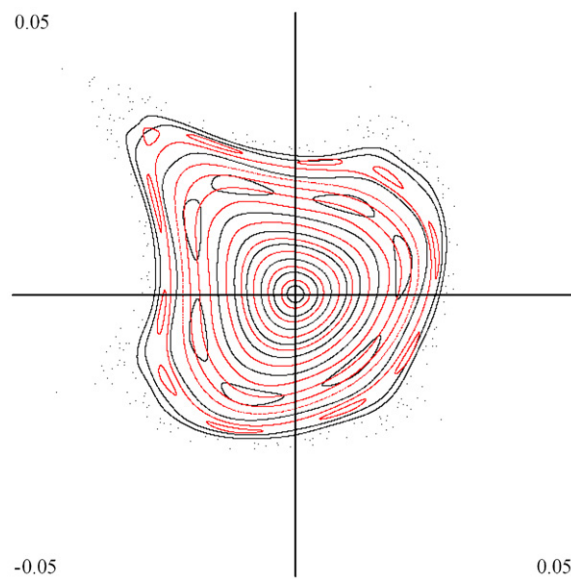
$$d_i \delta_j f_1 = \frac{\partial^2 f_1}{\partial q_i \partial p_j}. \quad (35)$$

Using this notation, the expression for  $G_2$  is simply

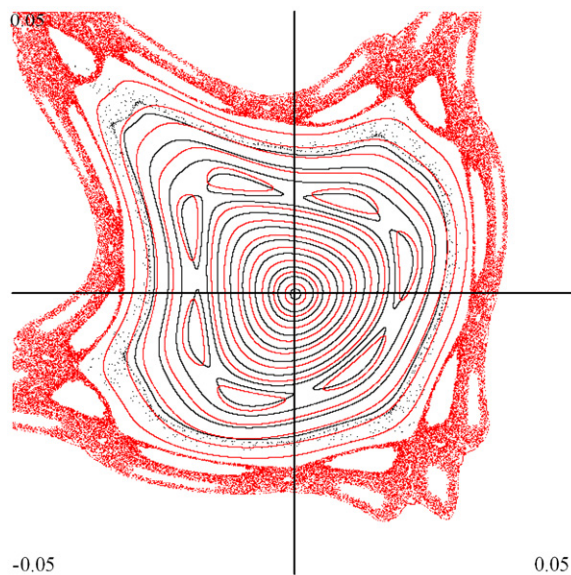
$$G_2 = f_2 - \frac{1}{2} d_i f_1 \delta_i f_1 \text{ (see equation (32))}. \quad (36)$$

In his hand written notes, Neri points out that the expression in rounded brackets came from  $G_2$  and the expression in square brackets came from  $G_3$ . I will leave it to the reader to

<sup>15</sup> I am grateful to the anonymous referee for pointing out that Dr David Douglas first derived the  $\nu = 4$  formulae. Most of the dates quoted here are from Dr Douglas' recollection. He also believed, falsely, that the idea of generating function tracking was mine. To the best of my knowledge, this was Professor Alex Dragt's idea. It was my idea to convert Taylor maps of TRANSPORT into Lie polynomial to be used as a MARYLIE input for the generating function tracking of MARYLIE. This happened when I was showed horrible nonsymplectic tracking by Dr Eugene P Colton of Los Alamos in 1983.



**Figure 6.** Phase space of using  $G_1$  in red and exact tracking in black.



**Figure 7.** Phase space of using  $G_1 + G_2$  in red and exact tracking in black.

compute  $G_3$ . I cannot guarantee that I copied correctly from Neri's notes. However I know that the implementation in MARYLIE 5.0 was correct<sup>16</sup> and therefore one could certainly reconstruct  $G_3$  and  $G_4$  from the software by contacting Alex Dragt.

<sup>16</sup> MARYLIE 5.0 did not contain fifth-order expressions for each magnet like COSY 5.0 of Berz; it contained only formulae to third order. Only concatenation formulae and other algorithms such as normal forms were performed to fifth order. The concatenation formulae were derived by Douglas and the normal form algorithm by Neri.

Finally, I must point out that Neri invented a strange graph calculus which is reminiscent of the Butcher [19] trees used in the theory of Runge–Kutta. I have been unable to verify if Neri was in anyway inspired by that work. His notes are filled with bizarre diagrams which I have been unable to decipher.

**3.3.3. Fitted generating functions.** Warnock, Ruth and Berg [20]<sup>17</sup> proposed the idea of fitted maps and fitted canonical transformations. The reader is also invited to look at [21] for fitted canonical transformations and applications to long-term stability bounds.

One can start with a tracking code, preferably a symplectic integrator, and simulate several trajectories. Let us assume that we have a canonical transformation  $a$  which very grossly normalizes the map of the code and thus we can assume that the resulting trajectories sit approximately on circles. Therefore we define action-angle variables as follows:

$$x_k - ip_k = \sqrt{2I_k} e^{i\Phi_k}. \quad (37)$$

By assumption, these actions are nearly constant for one iteration of the map. We then restrict ourselves to trajectories whose initial conditions sit on a product of annuli, more precisely,

$$\Phi_k^0 = 0 \quad \text{and} \quad I_k^0 \in [I_k^{\min}, I_k^{\max}], \quad k = 1, 2 \text{ or } 3. \quad (38)$$

The idea of Warnock was to represent the map of the tracking code in these annuli using a generating function of the form:

$$G(\Phi', I) = \sum_m g_m(I) e^{im \cdot \Phi'}. \quad (39)$$

Here the prime quantities are the values after one iteration. As usual, we get an implicit equation for the map:

$$\Phi = \Phi' + \frac{\partial G(\Phi', I)}{\partial I}, \quad I' = I + \frac{\partial G(\Phi', I)}{\partial \Phi'}. \quad (40)$$

The map itself can be written as

$$\Phi' = \Phi + R(\Phi, I), \quad I' = I + \Theta(\Phi, I). \quad (41)$$

We can extract the Fourier coefficients:

$$\begin{aligned} g_m(I) &= \frac{1}{(2\pi)^d i m_\alpha} \int_0^{2\pi} d\Phi' \frac{\partial G}{\partial \Phi'_\alpha} e^{-im \cdot \Phi'} \\ &= \frac{1}{(2\pi)^d i m_\alpha} \int_0^{2\pi} d\Phi' R_\alpha(\Phi(I, \Phi'), I) e^{-im \cdot \Phi'}. \end{aligned} \quad (42)$$

Since we do not have  $R$  as a function of  $\Phi'$ , we re-express this integral in terms of  $\Phi$ , i.e., in terms of functions of  $(\Phi, I)$  only:

$$g_m(I) = \frac{1}{(2\pi)^d i m_\alpha} \int_0^{2\pi} d\Phi R_\alpha e^{-im \cdot (\Phi + \Theta)} \det \left( 1 + \frac{\partial \Theta}{\partial \Phi} \right). \quad (43)$$

The zero mode must be handled differently, using the map for the angle. These integrals are discretized for evaluation using the tracking data. The Fourier coefficients are interpolated in the action variables using B-splines. The entire map is computed with a Newton search algorithm.

<sup>17</sup> Although my name appears as a co-author, my contribution was at best peripheral. I deserve credit for things never properly published such as proving that Ruth's fourth-order method is a splitting method and, conversely, deserves little credit for work properly published with my name attached to it. Perhaps this is a reflection of the fact that accelerator physics is not really an academic field. The powers ruling us do not care whether something is published or even correct: as long as the machine works!

These maps can be made, within the annuli, as precise as necessary if the underlying map is truly symplectic. One big advantage over the Taylor series map is clearly at large amplitude. It is possible to include effects which cannot easily be represented by a Taylor map such as the symplectic beam–beam kick of Hirata and Ruggiero in the weak–strong [22, 23] approximation.

Finally, I should say that the convergence [24] of these maps was proved by Drs R Warnock and J Ellison. They also discussed elsewhere [25] the issue of fitting maps to tracking data in Cartesian coordinates. Generating functions in action-angle variables do not work well if the ray is near the origin; in fact this map is often not partially invertible and therefore cannot be represented by a generating function.

### 3.4. Symplectification by jolts

Dr John Irwin while at the Central Design Group of the now defunct Super Conducting Super Collider [26] came up with an additional method for symplectifying a truncated Taylor series map. He first called it a ‘kick factorization’, but later Dragt and Abell [27] relabelled this method ‘jolt factorization’. First we define what Abell and Dragt called a jolt. It is a map of the form:

$$J = \exp(:g:)I = I + [g, I]. \quad (44)$$

For example, if the function  $g$  depends only on position, then equation (44) is true. In accelerators, these are called ‘kicks’ as in ‘kick codes’. However, it is also true if  $g$  depends only on the momenta. Moreover, if we have momentum in one degree of freedom and position in the other, equation (44) is again satisfied. Due to this generality, the term ‘jolt’ was adopted by Dragt and Abell. Here we will review a one-dimensional version of Irwin’s initial idea; the original paper of Irwin discussed the problem in two degrees of freedom.

Consider the following Lie map:

$$\mathcal{N} = \exp(:g_1^{N_0}:) \cdots \exp(:g_k^{N_0}:) \cdots \exp(:g_K^{N_0}:)\mathcal{R}. \quad (45)$$

In equation (45), all the  $g_k^{N_0}$  are jolts, specifically:

$$g_k^{N_0} = \sum_{n=1}^{N_0} \rho_k^n x_k^n, \quad \text{where } x_k = \cos \theta_k x + \sin \theta_k p. \quad (46)$$

The total number  $K$  of jolts necessary to represent a map to order  $N_0$  is  $N_0 + 1$ . In two degrees of freedom, it is determined by the subspace of polynomials generated by rotating  $x^k y^m$ . One chooses the subspace with the highest cardinality. The result is

$$\begin{aligned} K &= \frac{1}{4}(N_0 + 2)^2 & \text{if } N_0 \text{ is even,} \\ K &= \frac{1}{4}(N_0 + 1)(N_0 + 3) & \text{if } N_0 \text{ is odd.} \end{aligned} \quad (47)$$

Going back to the one degree of freedom case, we must select the angles  $\theta_k$ , following Irwin and later Abell, the best choice for these angles is:

$$\theta_k = \frac{k\pi}{K} \quad k = 0, K - 1, \quad \text{with } K \geq N_0 + 1. \quad (48)$$

In figure 8, we computed a jolt factorization equivalent to a third-order map using the minimum number of jolts. The reader will note that this minimum number of jolts is determined by the highest degree of the Lie polynomial as indicated by equation (48). This means that in one degree of freedom, the number of jolts is equal to the number of monomials in  $f_{N_0}$ , that is to say  $K = 5$  if  $N_0 = 4$ .

The result is not terribly good—in fact it is worse than the equivalent symplectification of figure 7. We may increase the number of jolts to  $K = 9$ . The results improve as shown in figure 9. We have always an excess of jolts compared to monomials; this over-determination is resolved by minimizing the usual norm defined as the sum of the square of the jolt strengths. Additionally, we can also ‘integrate’ the nonlinear part of the map. If the reader fast forwards to section 3.7, he will note that one can easily compute the  $(1/N)$ th power of a map and iterate it  $N$  times. Thus we can keep  $K$  constant<sup>18</sup> and apply any symplectic scheme to a map closer to the identity<sup>19</sup>.

One aspect of jolt factorization worth noting is that it is amenable to exact tracking without roundoff errors. Earn and Tremaine [28] pointed out that a map in the form of a jolt can be restricted to a grid and leaves this grid invariant. This map is exactly symplectic without roundoff errors in the sense that there exists a symplectic map which sends the grid onto itself and is otherwise defined elsewhere in phase space. The reader may wonder about the linear part of the map: this map can also be exactly factorized in terms of jolts. In fact, a symplectic integrator of the Ruth form does exactly that by splitting the expanded  $K_e$  of equation (6) along the so-called drift–kick split.

Tracking with a jolt factorization on a grid may have applications for very long-time simulations. One may start with a particle, in two or three degrees of freedom, in a region where the Taylor symplectified map is obviously adequate. We can then follow this map and see whether diffusion occurs without worrying about truncation errors. To my knowledge this was never done in accelerators.

I must admit that these discretized maps have not been tried in accelerators. Near the origin of the grid, a linear map is transformed into an utterly nonlinear map and thus diffusion is possible. These maps might introduce as many problems as they cure although the problems are perhaps mitigated by the fact that our system is not integrable and our investigations are not done near the origin. It was pointed out to me (by a reviewer) that Professor Franco Vivaldi has shown [29] that planar discretized rotations admit ‘an embedding into a dynamical system which is expanding with respect to a non-archimedean metric’. Vivaldi is also tied to our story through his studies [30] of the integrability of arithmetical maps. These maps were first studied by Edwin McMillan who was mentioned at the beginning of this paper in connection with the MURA project!

Finally I must point out that Dr Frank Schmidt of CERN, who is also the author of the code SIXTRACK, an early user of Taylor methods for perturbation theory, has done extensive numerical studies of the jolt factorization in collaboration with Abell; for example one can consult [31].

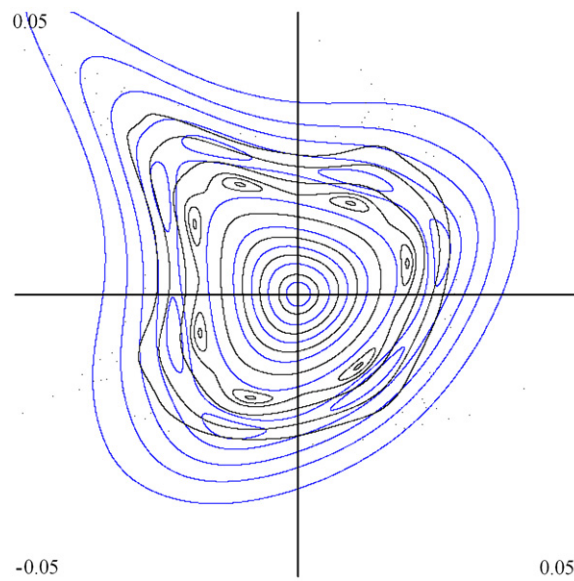
### 3.5. Symplectification by other methods

There are other ways to symplectify the Taylor map. One simple way, pointed to me first by Irwin but first published in a report in Maryland [32] by Gjaja, is simply to recognize that each monomial is exactly solvable. Let us consider an arbitrary monomial  $\alpha x^n p^m$ , it generates the following map:

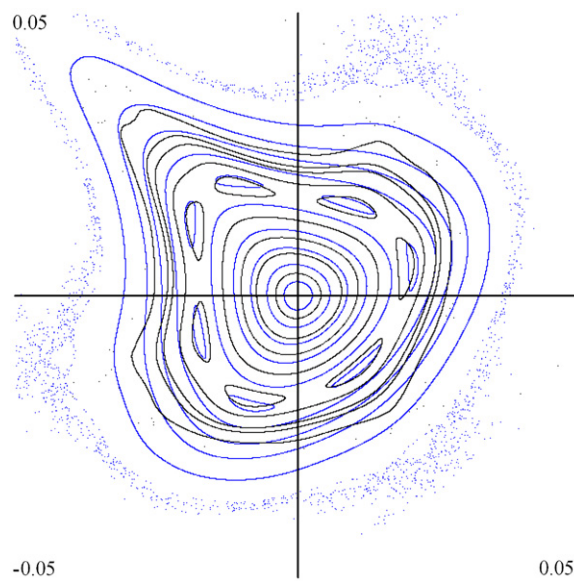
$$\begin{aligned} x(\alpha) &= \exp(:\alpha x^n p^m:)x = x(1 - \alpha(m-n)x^{n-1}p^{m-1})^{m/(m-n)}, \\ p(\alpha) &= \exp(:\alpha x^n p^m:)p = p(1 - \alpha(m-n)x^{n-1}p^{m-1})^{-n/(m-n)}. \end{aligned} \quad (49)$$

<sup>18</sup> Thanks to one of the reviewer for reminding me to state this often-used trick.

<sup>19</sup> One should not use the examples of this paper to conclude that symplectification with jolts is worse than the method of generating functions: a lot depend on details and implementation. For example, the Irwin method as well as Abell’s generalization depends on the norm one uses for a polynomial. Here, in all honesty, I did not use the norm suggested by Abell (see equation (63)).



**Figure 8.** Phase space of using  $K = 5$  jolts in blue and exact tracking in black.



**Figure 9.** Phase space of using  $K = 9$  jolts in blue and exact tracking in black.

For  $m = 0$  or  $n = 0$ , we get the usual jolts. For  $m = n$ , we can take the limit and regain an exponential map:

$$\begin{aligned} x(\alpha) &= \exp(:\alpha[xp]^n:)x = x \exp(-\alpha n[xp]^{n-1}), \\ p(\alpha) &= \exp(:\alpha[xp]^n:)p = p \exp(\alpha n[xp]^{n-1}). \end{aligned} \tag{50}$$

Thus, following Gjaja, we can take a map and rewrite it as a product of monomials. This can be done to an arbitrary order. One disease of the monomial factorization as well as the characteristic function methods are the existence of spurious poles and branch cuts in phase space. We already stated that ‘kick codes’ are very good approximations and therefore, if this is so, kick codes using the expanded Hamiltonian of equation (6) will always produce entire maps in phase space. There are a lot of perfectly inoffensive nonlinear maps which can develop spurious poles if one uses the monomial factorization. The reader will note that the jolt factorization does not have spurious poles.

This led Rangarajan [33] to look for yet another map factorization without spurious poles: the so-called polynomial factorization. We will see that it is actually another jolt factorization in disguise. For this reason, we will lump a summary of his ideas with those of Abell.

### 3.6. Generalizing jolts: Abell, Dragt and Rangarajan

The individual jolts in the Irwin factorization are created by rotating a simple kick (position jolts). More generally, in two degrees of freedom, one can think of acting on each individual kicks using members of a compact subgroup<sup>20</sup> of  $Sp(4)$ . Compactness is not absolutely necessary, but it allows us to define invariant norms. This was the idea of Abell and Dragt. In the original idea of Irwin, the subgroup used was  $SO(2) \otimes SO(2)$ .

*3.6.1. Polynomial maps viewed as jolts.* But before summarizing the work of Abell, one can describe the polynomial factorization of Rangarajan using the idea of extending the group acting on the jolts. Let us take an arbitrary kick in one degree of freedom where  $V$  is a polynomial:

$$\mathcal{K} = \exp(:V(q):). \quad (51)$$

This map can be modified by a simple scaling of the variables  $(q, p)$ :

$$A : \bar{q} = aq \quad \bar{p} = p/a. \quad (52)$$

We then act on the map  $\mathcal{K}$  using  $\mathcal{A}$ , the resulting map is still a jolt:

$$\mathcal{K}' = \mathcal{A} \exp(:V(q):) \mathcal{A}^{-1} = \exp(:V(aq):). \quad (53)$$

We can also introduce a  $p$ -dependent polynomial jolt  $g$  acting on the variable  $q$ :

$$B : \bar{q} = q + \frac{1}{a}g(p) \quad \bar{p} = p. \quad (54)$$

We scaled  $g$  by the constant  $a$  of equation (53) for reasons that will be apparent immediately: we can proceed to transform  $\mathcal{K}'$  by the Lie map  $\mathcal{B}$  associated with  $B$ :

$$\begin{aligned} \mathcal{K}'' &= \mathcal{B} \exp(:V(aq):) \mathcal{B}^{-1} = \exp(:V(aBq):) \\ &= \exp(:V[aq + g(p)]:). \end{aligned} \quad (55)$$

If both  $V$  and  $g$  are polynomials, then the effect of  $\mathcal{K}''$  on  $(q, p)$  is also to produce a polynomial map, i.e., a map akin to a pure jolt. To evaluate this map, we can use the factored expression of  $\mathcal{K}''$ ,

$$\mathcal{K}'' = \mathcal{B} \mathcal{A} \exp(:V(q):) \mathcal{A}^{-1} \mathcal{B}^{-1}, \quad (56)$$

<sup>20</sup> Warning to our friends in the mathematics community: physicists always use  $Sp(2n)$  as a short hand for  $Sp(2n, R)$ . It is not the unitary symplectic group  $U(2n) \cap Sp(2n, C)$ !

since every successive map is trivial. The result is

$$\bar{q} = q + \frac{1}{a}\{g(p) - g(p + aV'(aq + g(p)))\}, \quad \bar{p} = p + aV'(aq + g(p)). \quad (57)$$

Equation (57) is a polynomial if  $V$  and  $g$  are polynomials. We can go further, as Rangarajan did, and extend this result to a multidimensional phase space. Suppose that the constant  $a$  and well as the jolt  $g$  are functions of a set of mutually commuting (ignorable) variables in the other dimensions, then the result of equation (57) still holds for the plane  $(q, p)$ . In all generality, consider a canonical pair  $(\xi, \pi)$  and functions  $a$  and  $g$  that are functions of  $\xi$ . Then the effect of the map  $\mathcal{A}$  on  $\pi$  is

$$\bar{\pi} = \mathcal{A}\pi = \pi - qp \frac{\dot{a}}{a} \quad \text{where} \quad \dot{a} = \frac{da}{d\xi}. \quad (58)$$

The map  $\mathcal{B}$  can be computed as well:

$$\bar{\pi} = \mathcal{B}\pi = \pi + \dot{h}(p) \quad \text{where} \quad h = \frac{G}{a} \text{ and } G' = -g. \quad (59)$$

Again we can write the full map for  $\pi$  by composing all the transformations:

$$\bar{\pi} = \pi + \dot{h}(p) - \dot{h}(p + aV'[aq + g(p)]) + \frac{\dot{a}}{a}(aq + g(p))V'(aq + g(p)). \quad (60)$$

By expanding the function  $h$ , one can show that in equation (60) all the terms proportional to inverse powers of  $a$  vanish. This is the central result of Rangarajan: if  $V$ ,  $g$  and  $a$  are polynomials, then the final map is a polynomial map and therefore it is void of singularities. In this paper, I chose to make a connection with the jolt factorization even though it is not in Rangarajan's original work as far as I know. I did this because it ties everything neatly together but also because it makes it clear that this factorization can be put on a grid. As before, it suffices to factor  $\mathcal{A}$  in terms of jolts.

**3.6.2. Abell's generalization.** We have just seen that the polynomial factorization idea of Rangarajan is really a nonlinear extension of the group acting on the position or momentum jolts. Following Irwin's work, which used phase-space rotations  $SO(2) \times \cdots \times SO(2)$ , Abell and Dragt sought to generalize the linear group used by considering a more general compact subgroup of  $Sp(2n)$ . I will give here a small summary of Abell's work. The interested reader can consult Abell's thesis which is an excellent and a well-written document.

Abell starts with an arbitrary map of the form

$$\mathcal{M}_O = \exp(:f:), \quad (61)$$

where  $f$  is a homogeneous polynomial of degree  $O$ . Because of the Dragt–Finn factorization of an arbitrary map, we can treat each degree separately. Therefore from now on, we drop the reference to the degree. An arbitrary homogeneous polynomial  $f$  can be expanded in  $M$  monomials:

$$f = \sum c_k \vec{m}_k. \quad (62)$$

The monomials  $\vec{m}_k$  were normalized by Dragt and Abell as follows:

$$\vec{m}_k = \frac{z_1^{n_1} \cdots z_{2n}^{n_{2n}}}{\sqrt{n_1! \cdots n_{2n}!}}. \quad (63)$$

With this normalization, a standard dot product can be defined between polynomials

$$\langle f, g \rangle = \sum_k f_k g_k \quad \text{where} \quad f = \sum_k f_k \vec{m}_k \quad (64)$$

and it is invariant under the action of  $U(3)$ , the largest compact subgroup of  $Sp(6)$ . (See Dragt's book for a proof [34].)

Now, let us assume that we have a set of  $L$  linear Lie maps  $\mathcal{L}_l$  labelled by the index  $l$  and belonging to the compact part of  $Sp(2n)$ . Consider the complete set of all  $Q$  monomials of degree  $O$  which depend on position alone (kicks):

$$\vec{q}_j = \frac{x_1^{n_1} \cdots x_n^{n_n}}{\sqrt{n_1! \cdots n_n!}}. \quad (65)$$

Then one can construct a jolt as follows:

$$\vec{J}^{jl} = \mathcal{L}_l \vec{q}_j. \quad (66)$$

There are  $QL$  objects of this sort and they are all jolts. That is to say

$$\exp(\mathcal{L}_l \vec{q}_j) I = I + [\mathcal{L}_l \vec{q}_j, I]. \quad (67)$$

The numbers  $M$  and  $Q$  can easily be computed with  $M > Q$ . The number  $L$  is to some extent arbitrary. We saw in the case of the Irwin factorization ( $\mathcal{L}_l \in SO(2) \times \cdots \times SO(2)$ ) that the minimal value for  $L$  was determined by equation (47). Obviously, we must have  $LQ \geq M$ .

The idea, as with Irwin, is to rewrite  $f$  in terms of the jolts:

$$f = \sum c_k \vec{m}_k = \sum a_{jl} \vec{J}^{jl}. \quad (68)$$

Here the combination of indices  $jl$  should be thought as a single index describing  $LQ$  jolts. To get an equation for the jolt strengths, we compute the scalar product of with each monomial  $\vec{m}_k$ :

$$c_k = \sum a_{jl} \underbrace{\vec{m}_k \cdot \vec{J}^{jl}}_{\sigma_{jl}^k}. \quad (69)$$

Abell and Dragt introduce  $M$  sensitivity vectors  $\vec{\sigma}^k$  which live in a space of dimension  $LQ$ . Their analysis from this point on resembles closely a singular value decomposition. With this notation, the components  $c_k$  are obtained by simple scalar products of the sensitivity vectors and the unknown jolt strengths  $a_{jl}$ :

$$c_k = \vec{a} \cdot \vec{\sigma}^k. \quad (70)$$

It is clear from equation (70) that only the components of  $\vec{a}$  along the sensitivity vectors contribute to  $c_k$ . Therefore, Abell chose to write  $\vec{a}$  as a linear combination of the sensitivity 'vectors'.

$$\vec{a} = \sum_{k=1, M} \alpha_k \vec{\sigma}^k. \quad (71)$$

Substituting into equation (69)

$$c_k = \sum_{M, jl} \alpha_m \sigma_{jl}^m \sigma_{jl}^k = \sum_M \Gamma_{km} \alpha_m, \quad \vec{c} = \Gamma \vec{\alpha}. \quad (72)$$

Obviously, we can now solve for the coefficients  $\vec{\alpha}$  and therefore  $\vec{a}$ :

$$\begin{aligned} \vec{\alpha} &= \Gamma^{-1} \vec{c}, \\ \Rightarrow a_{jl} &= \sum_{m, k} \sigma_{jl}^m \Gamma_{mk}^{-1} c_k. \end{aligned} \quad (73)$$

At this stage we have done little more than Irwin. The solution in equation (73) minimizes the length of the component vector  $\vec{a}$ . This was also suggested by Irwin and achieved using Lagrange multipliers in Irwin's original work and in my own implementation. The only difference is that Abell chose a scaled monomial basis as shown in equation (63). The final results, given a set of jolts, depend on the scalar products selected for the components. It is not clear that Abell's choice, on pure dynamical ground, is the best choice, but it is the most convenient for the group theoretic arguments which constitute the core of Abell and Dragt's work.

I will now quote the salient results of Abell. First we note that a linear Lie map  $\mathcal{L}_j$  can be defined by its associated symplectic matrix  $R^j$ :

$$\mathcal{L}_j z_k = \sum_m R_{km}^j z_m. \quad (74)$$

According to Abell, when exploring the space of suitable matrices  $R^j$ , it suffices to consider only the part of  $R^j$  which is unitary. More precisely  $R^j$  can be factored into a modified Iwasawa form

$$R^j = \underbrace{\begin{pmatrix} F^j & 0 \\ G^j & F^{j-1} \end{pmatrix}}_{W^j} M(u^j), \quad (75)$$

where  $M(u)$  is a unitary symplectic  $2n \times 2n$  matrix made from the  $n \times n$  unitary matrix  $u$ :

$$M(u) = \begin{pmatrix} \text{Re}(u) & \text{Im}(u) \\ -\text{Im}(u) & \text{Re}(u) \end{pmatrix}. \quad (76)$$

Obviously in these expressions the phase-space vector  $\vec{z}$  is written as  $\vec{z} = (\vec{q}, \vec{p})$ . In equation (75), the matrix  $W^j$  only scrambles kicks (position jolts) amongst themselves and can be ignored in the search for a proper jolt basis according to Abell.

Thus Abell limits his search to  $U(n)$ . Abell then proves that one can further limit oneself to the coset  $U(n)/O(n)$ . In the case of the Irwin factorization, we limited ourselves to  $U(1)/O(1)$  in every phase plane.  $O(1)$  is just the trivial group containing 1 and  $-1$ . This restriction was achieved by limiting our rotation angles below  $\pi$ . Obviously a rotation of  $\pi$  on a kick at most changes its sign and therefore does not produce anything new.

Going back to equation (73), it is clear that the eigenvalues of  $\Gamma$  are crucial in an attempt to minimize the size of each jolt, i.e., the smallest eigenvalue of  $\Gamma$  must be as big as possible.

Naively we might think that a large set of evenly selected jolts from  $U(n)$  will indeed produce an optimal  $\Gamma$ . This has been verified semi-empirically by Abell. He also checked that if one selects the jolts randomly, particularly for a minimum set of jolts, then the smallest eigenvalue is truly small. Really this problem is tantamount to the proverbial needle in a haystack.

To conclude, Abell examined the continuous limit of  $\Gamma$  integrated over different subgroups of  $U(n)$ . He then determined the optimal smallest eigenvalue. Since we cannot use an infinite number of jolts on the computer, he then studied quadrature formulae (cubature for problems of higher dimensionality than  $U(1)$ ) which would approximate the integral over the subgroup as faithfully as possible. These studies require, amongst other things, the rewriting of the homogeneous polynomial space into subspaces carrying (irreducible) representation of  $SU(n)$ . This is a huge undertaking which goes well beyond the modest goals of this paper. The interested reader is encouraged to consult his thesis [27].

### 3.7. Symplectic restoration

Before getting into the modern era of symplectic integration started by Ruth, I would like to discuss why and how a Taylor map can be restored.

First, let us look at figure 10. I first generated a tenth-order map for our little one-dimensional cell. This map is extremely accurate and reproduces the results of exact integration all the way to the dynamic aperture. Then to actually produce this plot, I engaged in some ‘map molestation’: each of the nonlinear coefficients were randomly changed by approximately 5%. The various colours in figure 10 represent different trajectories made from 1000 iterations. I did not vary the linear part of the map so as to emphasize nonlinear effects. The big difference between this map and previous nonsymplectic maps is that it has a more or less constant level of accuracy when viewed as a Taylor series. Indeed by construction the coefficients of the Taylor series violate the symplectic condition uniformly by 5%. The process of restoration consists in computing the exact vector field of this nonsymplectic map, for example in the Dragt–Finn factorized form of equation (29) and in projecting this vector field on the space of symplectic vector fields.

For simplicity, let us imagine a map close to the identity which can be represented by a single generic Lie exponent:

$$N = \exp(F \cdot \nabla)I. \quad (77)$$

The vector field  $F$  is not symplectic. However, if we assume that it is close to a symplectic vector field, then

$$\exists f \quad \text{such that} \quad F \cdot \nabla I = F \approx [f, I] = -J \nabla f. \quad (78)$$

Using this equation, we can solve for  $f$ :

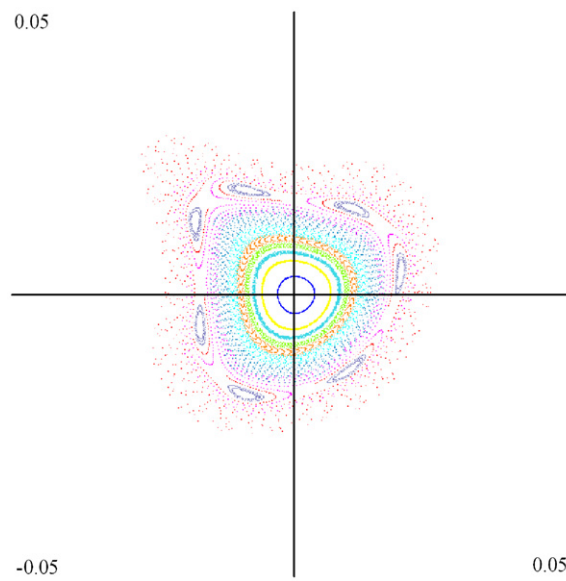
$$f = \int_0^z J F \cdot dz. \quad (79)$$

The uniqueness of the function  $f$  is assured if and only if the vector field  $F$  is symplectic, which is not in our case. In this calculation, I assumed that the symplectic condition is violated uniformly throughout phase space and thus I picked a ‘democratic path’ right on the straight line joining the origin and  $z$ :

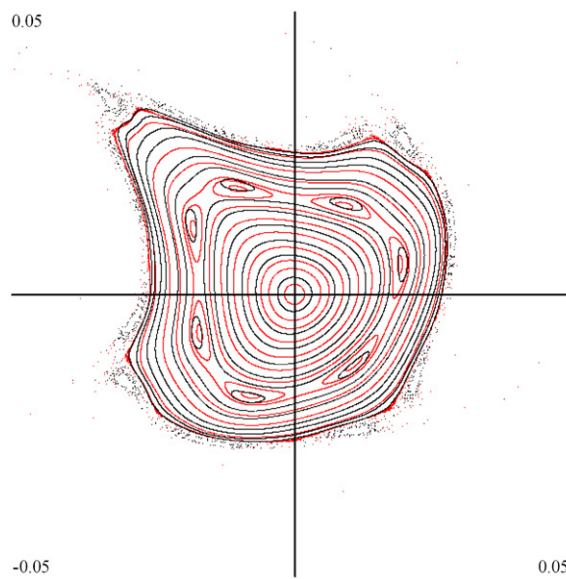
$$f = \int_0^1 J F(\alpha z) \cdot z d\alpha. \quad (80)$$

The result<sup>21</sup> of the restoration procedure is displayed in figure 11. The symplectic condition manifests itself, as seen in equation (27), by the equality of certain terms in the Taylor series expansion of the map. It can be shown that the choice of equation (80) for the restoration is equivalent to taking the average value of these terms and adding a term which is smaller than their difference. However, sometimes the source of symplectic violation is understood and a better path can be chosen. The author, in collaboration with Yiton Yan of SLAC, had such an experience in relation with the code TEAPOT and the late SSC ring project. We were trying to extract a high-order Taylor map from TEAPOT using automatic differentiation tools. We noted that the coefficients of the Taylor maps violated the symplectic condition around the eighth digit. The integrator TEAPOT was producing large floating point truncation errors in

<sup>21</sup> As I said at the beginning of this section, the original Taylor map was good enough (except for its violation of the symplectic condition due to my 5% ‘molestation’) and therefore nothing particular was done besides restoring the symplecticity of the Taylor map up to the order of truncation. Of course in general the other symplectification techniques (fabrication) described in this paper can be applied to this restored map.



**Figure 10.** Phase space of a map with a global error of 5%.



**Figure 11.** Phase space of a restored map in red and the original map in black.

the computation of the time of flight. The truncation errors in the transverse planes  $(x, y)$  were six to seven orders of magnitude smaller and so was the violation of the symplectic condition. Therefore, the high-order Taylor map, prior to tracking simulation, was restored by using a path of integration which minimized the usage of the  $(t, p_t)$  phase space. The integral of equation (80) was first performed along the  $(x, p_x, y, p_y)$  directions and then along the

$(t, p_t)$  directions. This insured that the restoration used mostly the pristine knowledge of the transverse plane.

*3.7.1. Why restoration?* Why do we want to use restoration? Obviously, the Taylor map from a nonsymplectic integrator will require restoration. In accelerators, this might be the case when non-trivial fringe fields are included. Even if we have an analytical representation of the fringe fields, they are not usually amenable to explicit symplectic integration.

There is another reason why restoration is sometimes an essential tool. On occasion, we must simulate the motion of a particle through a complex magnet whose magnetic field  $\mathbf{B}$  is known on a grid. The complex magnet is designed using a three-dimensional magnetic field solver and the field is computed at a selected number of grid points. There are several problems resulting from this type of representation:

1. Interpolation of  $\mathbf{B}$  leads to maps which are not differentiable beyond a certain level.
2. Symplectic integration is hard, unless we interpolate a vector potential.
3. And, of course, we would be restricted to implicit integrators.

So the trick I have personally used, inspired by the work of Berz and Makino (see Makino's thesis [36]), is to fit the three components of the magnetic field as a function of the transverse plane. Following Makino, we can use beamlets to represent the field:

$$B_k(x, y, z) = \sum_{i_x, i_y=1}^{N_x, N_y} \frac{B_k(i_x, i_y)}{S\pi^2} \exp\left(-\frac{(x - x_{i_x})^2}{\Delta x^2 S^2} - \frac{(y - y_{i_y})^2}{\Delta y^2 S^2}\right). \quad (81)$$

Here the  $z$  direction is the direction of integration. We do not need to fit in that direction. The quantities  $\Delta x$  and  $\Delta y$  are the horizontal and vertical grid spacings. The adjustable parameter  $S$  is normally set around 1. Actually Makino suggested a value of 1.8 for nearly constant fields and about 1.0 for varying fields.

Once we have smoothed the magnetic field in that manner, we can perform integration using noncanonical variables using  $z$  as the integration variable. The variable  $z$  can be a Cartesian length or an arc length.

Using the tools of automatic differentiation, we can compute the Taylor map for the full magnet. Since this map is from a field-free region to another field-free region, we can easily switch back to canonical variables—we bypassed completely the computation of a vector potential which would have required smooth functions in all three directions.

Still the resulting map will not be symplectic and it will require restoration. As a matter of fact, the difference between the restored coefficients and the old coefficients is a useful indication of the error in our procedure. Of course, violations of symplecticity come from two sources: the integrator is not symplectic and the magnetic field is not Maxwellian.

It seems to me that this is one of the most useful usages of symplectic restoration. It can also be coupled with a standard symplectic fabrication if necessary. It should also be added that this procedure is used on a single magnet map where accurate Taylor maps are most easily obtained.

Finally, one may compute a global  $\vec{A} = (A_x, A_y, 0)$  and use the theory germane to equation (106) or use an implicit integrator. In a wiggler or undulator, the average Hamiltonian is made, to a leading order, from the average of  $A_x^2 + A_y^2$ . This Hamiltonian, if expanded, is amenable to symplectic integration by Ruth-like splitting methods. This has been known for a long time (see Lloyd Smith's work [37]) but was also rediscovered by Dr Pascal Elleaume [38] using non-Hamiltonian variables.

*3.7.2. Generating function and vector field restoration.* At the beginning of this section on restoration, I mentioned that if one can compute a vector field for a map, then a Poisson bracket field can be extracted by a quadrature as shown in equation (79). This provides a method for symplectification. The same can be said with a generating function.

In the case of a purely nonlinear map, the process of extracting either a vector field or a generating function requires a finite number of steps. It is a recursive process, one first finds the cubic part of the generator using the quadratic part of the Taylor map. Then, one uses this to remove quadratic terms either exactly or approximately if the original map is not perfectly symplectic. Next we extract the quartic part of the generator from the cubic part and so forth and so on.

For a general map with linear terms, we can compute the vector field by an iterative process that converges for maps near the identity. It is tantamount to extracting the logarithm using the approximation for  $\log(1+x)$ . This iterative process requires us to compute

$$\begin{aligned} M &= \exp(F \cdot \nabla) I = \sum_{n=0}^{\infty} \frac{\{F \cdot \nabla\}^n}{n!} I \\ &= I + \sum_{n=1}^{\infty} \frac{\{F \cdot \nabla\}^n}{n!} I, \end{aligned} \quad (82)$$

which in turns necessitates an infinite number of evaluations if  $F$  has constant or linear terms. This can be done until convergence is reached. This is actually how the code COSY INFINITY [14] of Berz handles ideal magnets. The procedure can be sped up using a simple idea of Dragt, the so-called splitting and squaring idea (see [34]).

$$M = \left\{ \left\{ \left\{ \exp \left( \frac{F \cdot \nabla}{2^n} \right) I \right\}^2 \right\}^2 \dots \right\}^{2 \leftarrow n \text{ times}}. \quad (83)$$

The squaring is done by Taylor map concatenation. This procedure can speed up the exponential and is particularly useful in a code where such maps are constantly computed (modern matrix codes using automatic differentiation tools such as COSY INFINITY).

In the case of generating function symplectification, it suffices to perform a partial inversion of the nonsymplectic map. This can be done with a linear part as well. The generating function is then a simple integral similar to that of equation (79) except that the matrix  $J$  is replaced by the identity because the partially inverted map is exactly the gradient of a generating function. These techniques are quite standard.

I remind the reader that both techniques apply to linear and nonlinear maps.

### 3.8. Methods of restoration for purely linear maps

The discussion of section 3.7 applies to linear maps as well; for example the process of restoring the map extracted from fitted data includes its linear part. When dealing with purely linear maps, it is hard to imagine any process other than restoration. So in addition to the myriad of methods which might involve vector fields or characteristic functions, I will include here a few other ones.

*3.8.1. Gram–Schmidt symplectification.* This method, invented by the same Neri, and you guessed it, unpublished except in Dragt’s book [34], is used in the code MARYLIE. The idea is simple. We can view an arbitrary symplectic matrix  $M$  as a set of column vectors:

$$M = (\vec{m}^1 \dots \vec{m}^{2N}). \quad (84)$$

The symplectic condition reduces to conditions applied to the vectors:

$$\vec{m}^i \wedge \vec{m}^j = \vec{m}^{i\top} J \vec{m}^j = J_{ij}. \quad (85)$$

The idea consists in first ‘normalizing’ the vectors  $\vec{m}^1$  and  $\vec{m}^2$  using the above wedge product. The symplectified vectors  $\vec{s}^1$  and  $\vec{s}^2$  are given by

$$\vec{s}^{1,2} = \frac{\vec{m}^{1,2}}{|\vec{m}^1 \wedge \vec{m}^2|^{1/2}}. \quad (86)$$

Obviously we assumed here that the original vectors were nearly symplectic. As I said, this is truly a restoration procedure. We can continue with  $\vec{m}^3$  and  $\vec{m}^4$  by defining intermediate vectors which are not yet normalized but have a vanishing bracket with  $\vec{m}^1$  and  $\vec{m}^2$ . We introduce the intermediate vectors  $\vec{n}^3$  and  $\vec{n}^4$  and enforce a vanishing bracket with  $\vec{m}^1$  and  $\vec{m}^2$ :

$$\begin{aligned} \vec{n}^3 &= \vec{m}^3 + \alpha_1 \vec{m}^1 + \alpha_2 \vec{m}^2 \\ \vec{n}^4 &= \vec{m}^4 + \beta_1 \vec{m}^1 + \beta_2 \vec{m}^2 \\ \alpha_1 &= -\vec{m}^3 \wedge \vec{m}^1 \\ \alpha_2 &= \vec{m}^3 \wedge \vec{m}^2 \\ \Rightarrow \beta_1 &= -\vec{m}^4 \wedge \vec{m}^1 \\ \beta_2 &= \vec{m}^4 \wedge \vec{m}^2. \end{aligned} \quad (87)$$

The next step is simply to normalize the intermediate vectors:

$$\vec{s}^{3,4} = \frac{\vec{n}^{3,4}}{|\vec{n}^3 \wedge \vec{n}^4|^{1/2}}. \quad (88)$$

I leave it to the reader to extend this procedure to a higher dimension.

**3.8.2. Iterative symplectification.** Miguel Furman [39] invented at the SSC a method to symplectify a matrix by iteration. It is described in detail in Dragt’s book [34]. I summarize here Dragt’s exposition.

One can construct the following contraction map  $S(M)$  acting on a matrix  $M$ :

$$S(M) = (M J M^\top J^\top)^{-1/2} M. \quad (89)$$

One notes that if and only if  $M$  is symplectic, then  $M$  is a fixed point of  $S$ . Dragt shows in his book that a single application of  $S$  on a nearly symplectic map  $N$  projects its symplectic part defined through its polar symplectic decomposition. This decomposition states that a matrix  $N$  near the symplectic group can be factored as

$$M = Q R \quad (90)$$

where  $R$  is symplectic and  $Q$  obeys the relation

$$Q = J Q J^\top. \quad (91)$$

$R$  obeys the symplectic relation

$$R^{-1} = J R J^\top. \quad (92)$$

Furman’s idea was to approximate  $S$  to a leading order through an operator  $S_1$ :

$$S_1(M) = \frac{1}{2}(3I - M J M^\top J^\top) M. \quad (93)$$

This operator also leaves a symplectic matrix invariant and restores matrices whose deviation  $Q$  is close enough to the identity. It must be iterated until convergence is reached.

**3.8.3. Cayley symplectification.** Dragt also discusses a method which uses indirectly the exponential representation of the matrix  $M$  (if it exists). If and only if  $M$  is symplectic, it can be written as

$$M = (1 + T)(1 - T)^{-1}, \quad \text{where} \quad J^T T = \{J^T T\}^T. \quad (94)$$

The idea is to compute  $V = J^T T$  from  $M$

$$V = J^T T = (M + I)(M - I)^{-1} \quad (95)$$

and to symmetrize it into a new matrix  $W$ :

$$W = \frac{1}{2}(V + V^T). \quad (96)$$

Thus we can compute a symplectic matrix  $R$  using the symmetric  $W$ :

$$R = (I + JW)(I - JW)^{-1}. \quad (97)$$

Clearly in the case where  $M$  is symplectic, we have  $M = R$ , otherwise we have a symplectification scheme.

#### 4. Symplectic integration

As I have already mentioned, symplectic methods were present in accelerators before Ruth. The MURA precursors were aware of symplecticity and its importance. In the 1970s, the kick codes on the expanded Hamiltonian insured symplecticity naturally. Finally, the evil usage of matrix codes resulted in Dragt and collaborators developing the initial symplectification schemes on Taylor series using characteristic functions.

As we have already seen, the symplectification schemes do not guarantee a faithful map at large amplitudes. Therefore, for purposes of a short-term dynamic aperture, they are unreliable. Unfortunately, in science as in business, there are often conflicts of interest. Clearly it would be to the advantage of codes which stored the analytical Taylor series in memory to be able to handle short-term stability issues more reliably. Furthermore, one of the reasons for handling Taylor series was our ability to concatenate the maps into a single map for a large section of a machine. That concatenation, in turns, results in a degradation of the Taylor map as a tracking engine as we have seen. Any method to alleviate this problem was obviously welcome by proponents of Taylor series codes, including myself. And, not surprisingly, it led in the mind of some, to over-selling of the idea. Soon after my own entry into the field of accelerator physics, as a member of the SSC Central Design Group, I became convinced that indeed I had participated in the over-selling of the idea of symplectification. I became a ‘kick-code’ proponent.

I am not writing here a paper on the psychology of accelerator physicists, particularly since I was myself part of the delusion. However the initial push to sell symplectified maps has had perhaps some unintended bad consequences and some good ones. The good consequence is undoubtedly Ruth’s derivation of some symplectic integrators and the subsequent interest in our field and elsewhere. But, with that in mind, I will first describe the subsequent nasty fallout from the symplectification efforts.

As I said there is a theory of accelerator physics called the Courant–Snyder [6] theory. It is essentially a one degree of freedom linear Floquet theory to study the stability of a machine under linear errors and small nonlinear perturbations. Just as all philosophical works following Plato’s are footnotes to his work as Whitehead said, I will say that a great deal of the theory of perturbation in accelerator physics are footnotes to Courant–Snyder’s original paper.

One important ‘footnote’ is our ability to generalize the Courant–Snyder methods to more complex systems and to perform calculations in the cases when the re circulating machine is

represented by complex misaligned and mispowered models reflecting the unfortunate reality of accelerators. Thanks to the tools of automatic differentiation, introduced in our field by Berz, one can, from reliable integrators, extract Taylor maps on which an extended Courant–Snyder style Floquet theory can be used. This theory, initially proposed [40, 41] by Dragt and myself, is grounded theoretically and practically on the concept of finite  $s$  (our time-like variable) maps<sup>22</sup>. To make a long story short, throughout the years I have been advocating these theoretical and computational tools in conjunction with (preferably symplectic) integration. There are several reasons why these tools took a while to be fully accepted, but I am quite convinced that my association with the entire ‘symplectic fabrication’ led to confusion and to a certain scepticism. Despite my complete conversion to the symplectic integration methods, as far back as 1987 [43], my name is associated with ‘tracking with Taylor maps to the exclusion of anything else’. As recently as 2005, my most recent and complete implementation of symplectic schemes (the code PTC [44]) was listed on a web site of CERN (Centre Européen de Recherche Nucléaire), as a Taylor series code using truncated Taylor maps simulation although my main collaborators on PTC are CERN scientists!

So these are the cultural realities of my field of research. I will now describe reality as I see it following Ruth’s contribution, which like a seed, led to an immense growth of ideas and applications.

#### 4.1. Ruth’s integration

As I said, Ruth had strong views on the issue of symplectic fabrication. His views, shared by myself and most of the accelerator community these days, are that after one adopts a model, i.e., a Hamiltonian, then approximations in the integration variable are permissible while approximations in the phase-space coordinates are suspicious<sup>23</sup>. Therefore, we should use integrators as our main work horse. Besides the feeble kick codes, which were in effect second-order integrators using the matrix–kick split, little was known about higher order symplectic integration. Ruth, inspired by Laslett’s studies of nonlinear channels, went ahead and considered the following [5] Hamiltonian split:

$$H = A(p) + V(q). \quad (98)$$

To study symplectic integrators on Hamiltonians of the form of equation (98) is not without merit. We have already mentioned that the expanded Hamiltonian (6) has been used extensively in accelerators using the matrix–kick split. The drift–kick split was used in a code called SLIM [46, 47] developed by Professor Alex Chao for the computation of equilibrium beam sizes and equilibrium spin in electron machines.

So Ruth, as is well known to the geometric integration community in mathematics, published a paper [5] where he derived several symplectic integrators for equation (98). In that paper, the so-called more general<sup>24</sup> second- and third-order integrators attracted the attention of Neri circa 1985. Neri, the same Neri<sup>25</sup> of the symplectification formula in terms

<sup>22</sup> See also [42] from the ‘Italian school’.

<sup>23</sup> Talman and Peggs express this point of view in section 5.3 of [45]: ‘One popular, if suspect, procedure is to allow blemishes to develop while making approximations, but then to “paint over them” by re symplectifying.’ This is what I call fabrication in this paper.

<sup>24</sup> Less general methods in Ruth’s paper required the evaluation of derivatives of the force.

<sup>25</sup> Filippo Neri in the early 1980s became a postdoctoral fellow for Professors Alex Dragt and Bob Gluckstern at Maryland. He is presently a scientist at the Los Alamos National Laboratory. Although he has no published work on symplectic integration, he is a central character of this story in accelerator physics. I can still remember when he told me perhaps in 1985, ‘It is trivial using Lie methods to derive Ruth’s integrators.’

of generating functions, realized that if one were to write a general operator  $H$  acting on some linear space as

$$H = A + B, \quad (99)$$

then the exponential of that operator could be re-expressed approximately using the integration constants derived by Ruth for the Hamiltonian of equation (98). For example, if one has

$$e^{tH} = e^{tc_1 A} e^{td_1 B} e^{tc_2 A} e^{td_2 B} e^{tc_3 A} e^{td_3 B} e^{t^4 L(A, B)}, \quad (100)$$

then the coefficients of Ruth's original paper, given by

$$c_1 = \frac{7}{24}, \quad c_2 = \frac{3}{4}, \quad c_3 = -\frac{1}{24}, \quad d_1 = \frac{2}{3}, \quad d_2 = -\frac{2}{3}, \quad d_3 = 1, \quad (101)$$

are the solution of equation (100), that is to say, the remainder  $t^4 L(A, B)$  is an element of the commutator algebra of  $A$  and  $B$  at least proportional to  $t^4$ .

Ruth derived, during a sabbatical at CERN, a fourth-order symmetric integrator using generating function methods. He gave a talk on this topic at the Lawrence Berkeley laboratory around 1985 where he pointed out that his equation for the coefficients required him to solve a large equation containing hundreds of terms. Upon getting the solution numerically, he guessed that the cubic root of two entered in this number. Ruth then verified the analytical form of the coefficient. Personally, after talking to Neri, I derived the integrator using Lie formalism in a hand-written note (SSC-N-278, 1986). The end result was that Ruth's fourth-order integrator was really a splitting method:

$$\begin{aligned} \exp(tH) \approx & \exp\left(\frac{t}{2(1+\alpha)}A\right) \exp\left(\frac{t}{1+\alpha}B\right) \exp\left(\frac{\alpha t}{2(1+\alpha)}A\right) \exp\left(\frac{(\alpha-1)t}{1+\alpha}B\right) \\ & \times \exp\left(\frac{\alpha t}{2(1+\alpha)}A\right) \exp\left(\frac{t}{1+\alpha}B\right) \exp\left(\frac{t}{2(1+\alpha)}A\right), \end{aligned} \quad (102)$$

$$\alpha = 1 - 2^{1/3}.$$

This work is summarized in a paper of Ruth and myself [48]. The resulting integrator, later derived by hand independently by J Candy and W Rozmus [49] using Ruth's approach (*yes, by hand!*), was to be derived and generalized further<sup>26</sup> by H Yoshida [51], M Suzuki [52] and A D Bandrauk/H Shen [53]. We will come back to this later.

As of 1986, the following facts became clear to me and resulted in my full conversion from Taylor codes to integrators:

1. Ruth's integrators applied to any Hamiltonian system or any Lie system that can be split into two solvable parts.
2. The integrators of the so-called kick codes could be split along the matrix-kick or drift-kick axis. Ruth's method applied to both.
3. Ruth's method applied to the code TEAPOT and therefore to the full Hamiltonian used in Taylor codes such as TRANSPORT and MARYLIE.
4. Since we already knew that symmetrized Lie maps are quadratic for other reasons, it was also obvious that we had at least quadratic integrators for systems which required more splits.
5. Finally, since accelerator physicists are conservative people, it was important to point out that such an integrator would reproduce, if the number of steps and order were sufficiently high, the second-order matrices of TRANSPORT which were embedded in the CERN

<sup>26</sup> I was also contacted by the reviewer of Yoshida's paper, a specialist in Lie Poisson integration. He wanted to know if I was aware of Yoshida's generalization. It looked too trivially beautiful to have been overlooked. Well, I was not! In fact, in [50], sent for publication in 1990, I mentioned Yoshida post-mortem so to speak, i.e., the usefulness of my paper had died prior to publication.

code MAD [54], a code of biblical significance to accelerator physicists. This could be insured by the insertion of the automatic differentiation tools of Berz in the integrating code.

This list alone convinced me that a switch towards symplectic integration was best. Integration would allow us to study systems of arbitrary complexity and, with a map-based theory and automatic differentiation, we would lose nothing of the old Taylor maps tools embedded in the CERN code MAD. I expressed this opinion already in 1987 in [43] which already included a description of Ruth's fourth-order method as a splitting algorithm.

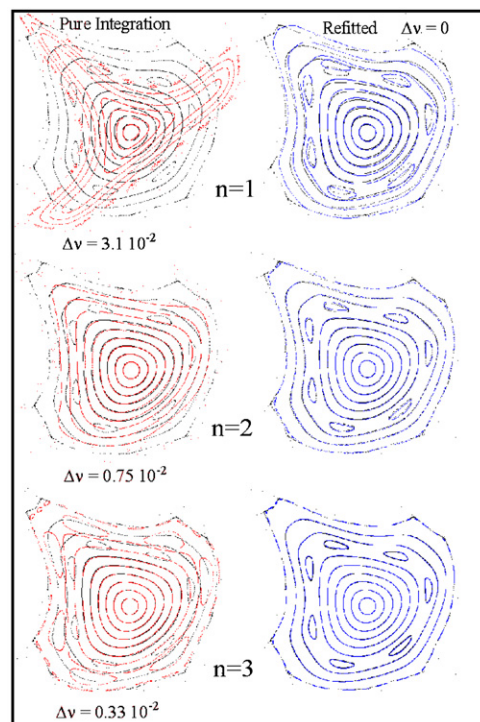
Unfortunately, the identification of map methods with symplectic fabrication and certainly vested interest in matrix codes insured that Ruth's idea would take longer to flourish in accelerators than elsewhere.

#### 4.2. Symplectic modellization or Talman's point of view

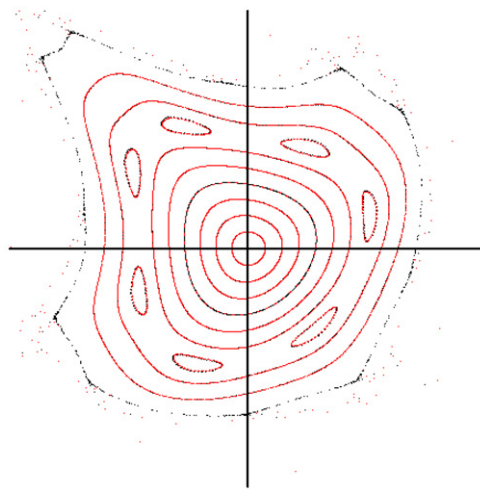
At this point it is tempting to introduce Yoshida's idea and come to the most modern point of view as fast as possible; however, let us follow the historical path and show why symplectic integrators are so efficient in our field using our little cell example once more.

Let us look at a few plots of our little cell system using the drift-kick split. In figure 12, I used one, two and three steps of the quadratic leap frog method (drift-kick-drift). On the left side, we note pure integration. The quantity  $\Delta\nu$  is the amount by which the so-called tune is miscalculated by pure integration. The tune is related to the eigenvalue of the system by the formula  $\lambda = \exp(i2\pi\nu)$ . Obviously the case with a single step,  $n = 1$ , is unacceptable. The case  $n = 2$  is not that great either. First of all there are no islands and the dynamic aperture is slightly lower. Finally, the results for  $n = 3$  are quite similar already to the exact results in black. To understand the right-hand side of figure 12, we must go back to the venerated matrix-kick-type split used in equation (6). Kick codes of the matrix-kick type commonly used a single kick per sextupole and the exact matrix for the quadratic part of the Hamiltonian. It is instructive to compare this with the exact result. This is done in figure 13. One sees that the matrix-kick split, with a single kick per nonlinear magnet, does nearly a perfect job. The necessity to get the linear part 'right' was used as an argument against the drift-kick-split (used in Chao's SLIM code for example) and to some extent gave another reason for Ruth to search for a higher order method. However, since it seems that the linear properties of the system dominate it so totally, it led Talman, in connection with his code TEAPOT, to take a more engineering point of view. Rather than saying that we integrate a Hamiltonian, it is sufficient to say that we have a model for the machine. This model should be tuned so as to reproduce the properties of the machine which are measurable and known empirically (and theoretically in some cases) to affect the machine operation.

It turns out that linear properties of a machine are crucial and are measurable. When an accelerator is turned on for the first time, even beams of very low current which are not affected by considerations outside the scope of our discussion, will most likely be lost. The magnets' exact fields, the misalignments of the magnets, even on the micron level, will produce initially an unusable machine. Talman's idea was to treat the integrator as an empirical model and to fit this model to the known and desired properties of the machine as we do when we start the operation of a real machine; this is obviously automatically so in the old kick codes with the matrix-kick split. On the right-hand side of figure 12, I did exactly this for our little cell example. The results for even  $n = 1$  are remarkable and much better than the results of 'symplectic Taylor fabrication'. In my view both Ruth and Talman were vindicated. Talman insisted to give even a further degree of physical realism to his model, for

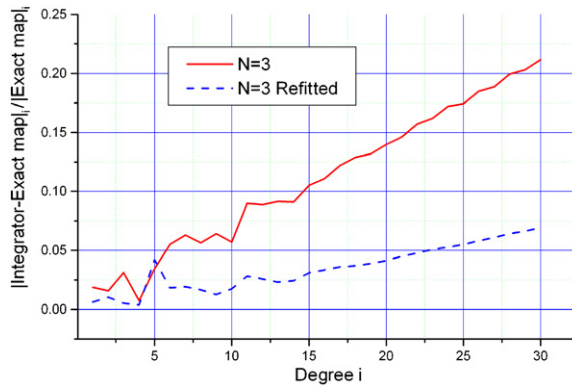


**Figure 12.** Second-order integration in red or blue, exact results in black.



**Figure 13.** Matrix-kick split in red.

example, rejecting backward drifts required in higher order formulae such as Ruth's fourth order. This particular aspect of Talman's TEAPOT, the insistence on being able to visualize each function evaluation of the integrator as something perhaps physically realizable, is in



**Figure 14.** Norm of each degree.

my own experience unnecessary. As long as we avoid transverse approximations, i.e. use integrators, and combine that with refitting for additional efficiency, we will produce more reliable simulations with integrators than with Taylor codes<sup>27</sup>.

#### 4.3. Symplectic integration is restoration

I have mentioned already that symplectic integration is a form of restoration. In figure 14 we see the slow increase of the error in the Taylor coefficients as a function of order. Standard integration and refitted integration are compared. Here the crucial parameter is the ‘phase advance’ per cell and the total phase advance of the ring. For example, in the old SSC for which the code TEAPOT was designed, there were arcs made with numerous cells whose phase advance was either  $60^\circ$  or  $90^\circ$ . This meant that the eigenvalues of the linear map of each cell were either  $e^{i\pi/3}$  or  $e^{i\pi/2}$ . In such a system, it is easy to show by considering the nonlinear map for each cell that certain non linearities will be additive over the full arc and many others will exactly cancel out. Of course, this requires the cell to be set exactly<sup>28</sup> at  $60^\circ$  or  $90^\circ$ . Talman, using a single step of integration per arc quadrupole and bending magnets, adjusted their values to precisely preserve the focusing properties of each cell. If I were to try on an SSC arc the calculations displayed on the left-hand side of figure 14, the difference between the two curves would be enormous. This led many accelerator physicists to dismiss incorrectly the drift–kick split before Talman’s introduction of TEAPOT. Obviously, Talman’s refitting of the model coupled with higher order techniques greatly enhanced the respectability of symplectic integration in accelerator physics. The reader will also note that the extraction of Taylor maps from integrators, which is necessary for the creation of figure 14, became simultaneously feasible around 1987 due to Berz’s work. In my view, these advances set the stage for the replacement of Taylor codes by integrator codes. In any event, an integrator code can compute a Taylor map and therefore nothing is lost.

<sup>27</sup> I did not explain in detail the refitting procedure. In the example of this section, the motion was integrated using the drift–kick method. Thus the refitting procedure involved changing the quadrupole strength and bringing back the tune to its original value, i.e.,  $\Delta\nu = 0$ . Obviously if the matrix–kick split is used no refitting is needed: this is clearly illustrated in figure 13.

<sup>28</sup> More precisely it requires the absence of systematic errors in the phase for this to be approximately true.

#### 4.4. Yoshida and implications

It is clear that the theoretical improvements facilitated the derivation of new integrators. For example once Neri realized that Ruth's integrators were splitting methods, I verified easily that the unpublished fourth-order Ruth integrator was indeed a splitting method and I derived new sixth-order integrators [50] including integrators customized for the expanded Hamiltonian of equation (6). But, as we have just seen, the Talman modelling view makes it clear that we can quite often survive with relatively low-order integrators. Therefore, in my view, the major impact of introducing Lie methods to symplectic integration was to extend the work of Ruth beyond what Ruth himself would have thought possible. In fact, we will see that the last problem posed by Ruth in his original paper [5], which is the inclusion of an arbitrary vector potential within the expanded Hamiltonian framework (equation (6)), is completely solved using Yoshida's point of view.

In all of this, I proselytized inspired by Ruth's idea but above all using the various extensions of the theory made possible by the people who followed. The greatest extension is Yoshida's realization that Ruth's fourth-order integrator is actually made of three steps of a second-order time-reversible integrator. Therefore, if one can write a second-order time-reversible integrator, one can automatically write a fourth-order one. In fact, a  $(2n+2)$ th-order integrator is obtainable trivially from a symmetrized product of  $2n$ th-order integrators.

It was already known to us that symmetrized products are time-reversible second-order maps. In fact we design the interaction region of colliders so that magnets appear in a symmetric arrangement around the collision point. So this aspect of Yoshida's method was an actual magnet configuration for us! We also knew, as I will later discuss that splitting methods can involve exact solutions of non-trivial unperturbed [55] problems. Moreover, we were aware from the very beginning that changing the number of integration steps would break the symplecticity. Finally, we knew of the existence of the unpublished Ruth fourth-order integrator, circulating semi-secretly since 1984 amongst practitioners of our trade. Given all these facts, and in particular that I, a Lie method user, was personally aware of all of this early on, the discovery of Yoshida was particularly embarrassing<sup>29</sup> for myself. But this is how the cookie crumbles.

Before giving an example, let us summarize Yoshida's simple but yet profound discovery. Let us assume that we have a map  $\mathcal{M}$  and that it is approximated by a map  $\mathcal{S}$  that has the following properties:

$$\mathcal{M}(\Delta t) = \mathcal{S}_2(\Delta t) + O(\Delta t^3), \quad \mathcal{S}_2(-\Delta t) = \mathcal{S}_2^{-1}(\Delta t). \quad (103)$$

For example, if  $\mathcal{M}$  is generated by a Hamiltonian  $H = \sum_{i=1}^M H_i$ , then a suitable  $\mathcal{S}_2$  can be

$$\mathcal{S}_2(\Delta t) = e^{(-\frac{\Delta t}{2}:H_1:)} e^{(-\frac{\Delta t}{2}:H_2:)} \dots e^{(-\frac{\Delta t}{2}:H_2:)} e^{(-\frac{\Delta t}{2}:H_1:)}. \quad (104)$$

Then, a fourth-order integrator can be constructed using  $\mathcal{S}_2$ :

$$\begin{aligned} \mathcal{S}_4(\Delta t) &= \mathcal{S}_2(x_1 \Delta t) \mathcal{S}_2(x_2 \Delta t) \mathcal{S}_2(x_1 \Delta t), \quad \text{where} \quad x_1 = \frac{1}{2 - 2^{1/3}} \\ \text{and} \quad x_2 &= \frac{-2^{1/3}}{2 - 2^{1/3}}. \end{aligned} \quad (105)$$

This integrator applied to a system split into exactly two solvable parts reproduces Ruth's fourth-order integrator as reinterpreted by Neri and Forest. Yoshida's procedure can, in addition, produce arbitrary-order integrators. Indeed we can recursively apply equation (105)

<sup>29</sup> When I mentioned to Ruth that this discovery of Yoshida made us look dumb, he responded quite correctly: 'Speak for yourself!'

to get integrators of increasing order. This does not necessarily lead to the most compact integrators. Yoshida derived special sixth- and eight-order integrators which use fewer time steps. For accelerators, sixth order is about as far as one will ever need especially in the light of Talman's modelling concept.

As an example of the importance of Yoshida's construction, we will address the last unsolved problem of Ruth: the inclusion [56] of a vector potential in equation (6). The most general magneto-static Hamiltonian (for small angle) has the form,

$$K = -p_t + \frac{(p_x - a_x)^2 + (p_y - a_y)^2}{2(1 + p_t)} - \frac{x p_t}{\rho} + \frac{x^2}{2\rho^2} + V_2 + V_{\geq 3}, \quad (106)$$

where  $a_x$ ,  $a_y$  and  $V$  are functions of the integration variable  $s$  as well as  $x$ ,  $y$ . The first problem resides in the  $s$  dependence—our variable of integration. This is easily handled by extending the phase space:

$$K_\sigma = p_s - p_t + \frac{(p_x - a_x)^2 + (p_y - a_y)^2}{2(1 + p_t)} - \frac{x p_t}{\rho} + \frac{x^2}{2\rho^2} + V_2 + V_{\geq 3}, \quad \frac{ds}{d\sigma} = 1. \quad (107)$$

Next, we examine the vector potential term

$$K_x = \frac{(p_x - a_x)^2}{2(1 + p_t)}, \quad (108)$$

which we recognize as exactly solvable in the extended phase-space formalism. The formal solution is

$$\begin{aligned} \mathcal{M}_x(\Delta\sigma) &= \exp\left(-\Delta\sigma : \frac{(p_x - a_x)^2}{2(1 + p_t)} : \right) \\ &= \exp\left(: \int^x a_x dx : \right) \exp\left(-\Delta\sigma : \frac{p_x^2}{2(1 + p_t)} : \right) \exp\left(- : \int^x a_x dx : \right). \end{aligned} \quad (109)$$

In equation (110), we recognize a horizontal drift transformed by a jolt given by the integral of  $a_x$  with respect to  $x$ . A map  $\mathcal{M}_y$  can be similarly constructed in the  $y$  plane. Thus we have the following split for the full Hamiltonian:

$$K_\sigma = \underbrace{p_s - p_t}_1 + \underbrace{\frac{(p_x - a_x)^2}{2(1 + p_t)}}_2 + \underbrace{\frac{(p_y - a_y)^2}{2(1 + p_t)}}_3 + \underbrace{-\frac{x p_t}{\rho} + \frac{x^2}{2\rho^2} + V_{\geq 2}}_4. \quad (110)$$

Using, the symmetrized expression of equation (104) and the above split, we can immediately write the beginning of a Yoshida hierarchy of integrators. This was used in the code PTC [44] and introduced by Sagan in the code BMAD [57] of Cornell.

Another extension in accelerators [58] uses the Poincaré implicit quadratic generating function:

$$\begin{aligned} q^f &= q + \Delta t \frac{\partial H}{\partial x_2} \left( \frac{q + q^f}{2}, \frac{p + p^f}{2}, t + \frac{\Delta t}{2} \right), \\ p^f &= p - \Delta t \frac{\partial H}{\partial x_1} \left( \frac{q + q^f}{2}, \frac{p + p^f}{2}, t + \frac{\Delta t}{2} \right). \end{aligned} \quad (111)$$

This is being used by J S Berg of Brookhaven National Laboratory [59] for magnets with a large aperture where the fringe fields are important. Obviously, it is a quadratic reversible implicit method and therefore Yoshida's hierarchy applies. A similar canonical transformation has been advocated by Berz and Erdelyi in [18] as a mean to symplectify the one-turn Taylor map.

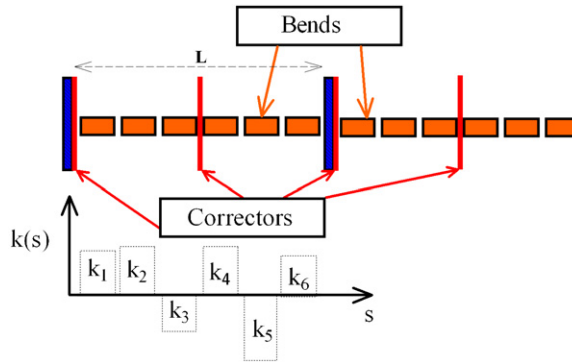


Figure 15. SSC cell with correctors.

#### 4.5. Biased integrators and correcting schemes

The high-order integrators of Yoshida and Ruth treat equally all terms of the splitting method. If the Hamiltonian is split as  $H = H_1 + H_2 + H_3$ , then the fourth-order Yoshida gives us

$$S_4(\Delta t) = \exp(-\Delta t : H : + O(\Delta t^5)). \quad (112)$$

However, the way accelerators are made the main contribution to the dynamics comes from quadratic term of  $H$ , let us say  $H_1$  of the above split. Therefore one concludes that it is more important to get the terms which are of higher order in  $H_1$  correctly while tolerating second-order integration as far as the nonlinear corrections are concerned.

This is essentially the approach in PTC and SAD. To some extent it is also the approach in the original TEAPOT where special integrators were concocted for the powerful interaction region quadrupoles. In PTC, I use a Simpson and Bode rule integrator for the splits of equations (131) and (132). In the case of the Simpson rule method, a single time step is given by

$$\begin{aligned} \mathcal{B}_4 = & \mathcal{M}_3 \left( \frac{\Delta t}{6} \right) \mathcal{M}_2 \left( \frac{\Delta t}{6} \right) \mathcal{M}_1 \left( \frac{\Delta t}{2} \right) \mathcal{M}_3 \left( \frac{\Delta t}{3} \right) \\ & \times \mathcal{M}_2 \left( \frac{2\Delta t}{3} \right) \mathcal{M}_3 \left( \frac{\Delta t}{3} \right) \mathcal{M}_1 \left( \frac{\Delta t}{2} \right) \mathcal{M}_2 \left( \frac{\Delta t}{6} \right) \mathcal{M}_3 \left( \frac{\Delta t}{6} \right). \end{aligned} \quad (113)$$

The ‘nonlinear terms’ produced by  $H_2$  and  $H_3$  appear in the proportion (1, 4, 1) and thus I call this method a Simpson’s rule method. In the jargon of accelerator physicists, the effect of the phase advance ( $H_1$ ) on the nonlinear terms is handled correctly to fourth order. Otherwise it is just a second-order method by virtue of symmetrization. One can pursue this further with a distribution in the proportion (7, 32, 12, 32, 7). I call this a Bode integrator.

In accelerator physics these biased integrators originated as correction schemes for the half cell of the SSC. A need emerged to correct a constant multipole component due to eddy currents in the six superconducting bends of the half cells of the gigantic machine. Since these effects are constant, it led to a (1, 4, 1) setting of the correctors. It did not take a huge amount of intelligence on my part to reverse the concept and make this an integrator. Neuffer came up with an idea of the (1, 4, 1) corrector; see [60] for example.

Later, we faced the potential need to correct different errors in each magnet as shown in figure 15. This led us to a more general Neuffer correction scheme for which the ratio (1, 4, 1) had to be replaced by a ratio depending on the actual measured magnetic field error. We present the derivation of this general correction scheme because it leads directly to a simulation code,

the ‘Ritson code’ or SSCTRK [61], which was tailored to the weak focusing properties of the bends of the SSC half cell. Therefore, it also falls within the domain of symplectic methods.

*4.5.1. The Neuffer corrector and the Ritson approximation.* This generalized Simpson idea of Neuffer and Forest was used by Ritson in the days of the SSC to speed up tracking [61] by lumping all the nonlinearity of a single half cell at two positions. This made sense because the six bends of the half cell provided little focusing, i.e., little phase advance. Perversely the idea originated as a corrector scheme proposed by Neuffer [60], Peterson and Forest. If one can mimic the nonlinearity of six consecutive magnets using two Simpson-like positions, why not reversing the polarity and approximately correcting these nonlinearities by installing correctors at the beginning and in the middle of the half cell! So geometric integration à la Ritson and actual correction schemes of Neuffer *et al* were part of a single theoretical framework. To illustrate the theory, we will neglect the small drift space between the bends and we assume that the bends are drift spaces perturbed by a potential  $V$ . So the Hamiltonian for the half cell is given by

$$H = D + \sum_{i=1}^6 k_i(s) V. \quad (114)$$

The original idea of Neuffer and Forest was to install thin lens correctors at the quadrupole location in order to cancel the first-order Lie polynomial  $f$  of equation (116) as much as possible. Conversely, Ritson’s idea was to use the strength of the correctors to simulate an uncorrected cell with the help of kicks at the quadrupole and at the half cell location. This could speed up the tracking of the gargantuan SSC lattice considerably.

Thus, consider the modified Hamiltonian

$$H = D + \left\{ L \left( c_1 \delta(s) + c_2 \delta \left( s - \frac{L}{2} \right) + c_3 \delta(s - L) \right) + \sum_{i=1}^6 k_i(s) \right\} V. \quad (115)$$

The map going through the six bends of the half cell can be computed to first order in the potential  $V$ :

$$\mathcal{M} = \exp(-:f:) \exp(-L:D:),$$

$$f = \int_0^L ds e^{-s:D:} \left\{ L \left( c_1 \delta(s) + c_2 \delta \left( s - \frac{L}{2} \right) + c_3 \delta(s - L) \right) + \sum_{i=1}^6 k_i(s) \right\} V. \quad (116)$$

The equation for  $f$  in (116) is expanded to second order in  $s$  and each term is set to zero. We obtain three equations:

$$\begin{aligned} c_1 + c_2 + c_3 &= \pm \frac{1}{6} \sum_{i=1}^6 k_i = \langle k \rangle, & \frac{1}{2} c_2 + c_3 &= \pm \sum_{i=1}^6 \left\{ \frac{2i-1}{72} \right\} k_i = \langle sk(s) \rangle, \\ \frac{1}{4} c_2 + c_3 &= \pm \sum_{i=1}^6 \left\{ \frac{3i^2 - 3i + 1}{648} \right\} k_i = \langle s^2 k(s) \rangle. \end{aligned} \quad (117)$$

The solution is

$$\begin{aligned} c_1 &= \mp \{ 2 \langle s^2 k(s) \rangle - 3 \langle sk(s) \rangle + \langle k \rangle \}, \\ c_2 &= \mp 4 \{ \langle sk(s) \rangle - \langle s^2 k(s) \rangle \}, \\ c_3 &= \mp \{ 2 \langle s^2 k(s) \rangle - \langle sk(s) \rangle \}. \end{aligned} \quad (118)$$

The upper sign refers to Ritson style integration and the lower sign to the Neuffer–Forest correction schemes.

4.5.2. *The true biased methods.* Obviously, for a constant  $k$  across a single magnet we regain the biased Simpson scheme of the beginning of this section. We get for the various moments

$$\langle k \rangle = k \quad \langle sk(s) \rangle = k/2 \quad \langle s^2 k(s) \rangle = k/3 \quad (119)$$

and, substituting in equation (118), we obtain

$$c_1 = \frac{1}{6}k, \quad c_2 = \frac{2}{3}k, \quad c_3 = \frac{1}{6}k. \quad (120)$$

The particular historical context and certain idiosyncrasies of our field kept another solution of the biased integration out of view. Correction schemes do not permit us to imagine magnets of variable sizes while it is perfectly acceptable for an integration scheme, i.e., we can cut the linear part of the magnet as we please. If we choose two ‘correctors’ placed symmetrically at a position  $s_c$  and  $L - s_c$  respectively, we must then choose  $s_c$  so as to zero the moment expansion of  $f$  which is now given by

$$f = \int_0^L ds e^{-s:D} \left\{ L \left( \frac{1}{2} \delta(s - s_c) + \frac{1}{2} \delta(s - L + s_c) \right) - k \right\} V. \quad (121)$$

Re-expressing  $s_c$  as  $L\xi_c$ , we get an equation for  $\xi_c$  using the quadratic moment since the lower moments automatically vanish thanks to our symmetrical choice of strength and positions:

$$\frac{1}{2}\xi_c^2 + \frac{1}{2}(1 - \xi_c^2)^2 = \frac{1}{3}. \quad (122)$$

The solution is given by

$$\xi_c = \frac{3 - \sqrt{3}}{6}. \quad (123)$$

This solution was apparently derived by myself and Neuffer in our original paper [60] on correction schemes. I say ‘apparently’ because it appears only in a figure where it is called a ‘two-point Gaussian-quadrature-like correction’. I assumed that, obsessed as we were with correction schemes for the SSC half cell, Neuffer barely consented that I quote this unphysical solution without proof.

Going to the Simpson- and Bode-type integrators, we note that they have two properties: they require equal steps for the Hamiltonian  $H_1$  and the kicks for the remaining terms of the splits are all positive. We can ask if these two properties carry over to higher order methods. The answer is ‘no’. The next method still has that property. The magnet is divided into six equal parts as far as  $H_1$  is concerned. The kicks are given by

$$\frac{c_i}{k} = \left( \frac{41}{840}, \frac{9}{35}, \frac{9}{280}, \frac{34}{105}, \frac{9}{280}, \frac{9}{35}, \frac{41}{840} \right). \quad (124)$$

This method is 8th order in the effect of  $H_1$  on the remaining terms. However, if we attempt a tenth-order method, then negative coefficients  $c_i$ ’s are required. As I said before, this type of method was unpleasant to Talman for reasons having to do more with unjustified prejudices<sup>30</sup> than real physics or mathematics.

It turns out that this entire theory was later derived and explored further independently by McLachlan [62]. In fact I had McLachlan’s note in my collection of preprints since 1994 but, due to difference in language and culture, I had not realized when I first received it that it was on a topic so closely related. McLachlan talks about ‘Composition Methods in the Presence of Small Parameters’ while we titled our work ‘A General Formalism for Quasi-Local Correction

<sup>30</sup> In all honesty, if Talman had not had his full baggage of prejudices, TEAPOT would never have seen the light of day simply because slightly retouched kick codes with the expanded Hamiltonian would have been enough for the SSC. Indeed LHC tracking at CERN was entirely done with codes without the square root Hamiltonian. And I perhaps would have lost all interest in symplectic integrators and would not be writing this paper!

of Multipole Distortions in Periodic Transport Systems’—in fact, I first used this ‘biased’ epithet at a later date in my lectures and my book [63].

Later, Laskar and Robutel pushed the idea of a ‘biased’ integrators using only positive steps in a paper titled ‘High order symplectic integrators for perturbed Hamiltonian systems.’ They purposely avoid negative coefficients for both  $H_1$  and the remaining  $H_{\geq 2}$ . Of course while these methods require less steps than our biased methods, they necessitate steps of varying sizes for  $H_1$  for methods beyond the fourth order. At the fourth order, Laskar and McLachlan reproduce the Simpson and Gaussian quadrature methods.

In that paper, it is written that ‘According to Yoshida, (1990), the search of symplectic integrators using Lie formalism was introduced by Neri (1988).’ It is certainly true that Neri was the first person to state that Ruth’s methods were Lie methods (splitting methods). However it certainly predates 1988. In [43], I already referred to Ruth’s third-order method as a Lie method which I had personally derived in an SSC note in 1986. To the best of my knowledge, Neri’s idea was already circulating in 1985. Finally, the idea of generating integrators which treat the feed-down of  $H_1$  with a greater accuracy goes back to Talman<sup>31</sup>, Neuffer and myself circa 1986. It remains true that McLachlan and Laskar explored the ideas with far greater depth than any one of us did in accelerators.

## 5. Actual code using symplectic methods

On the one hand, by the end of 1986, I had come to the conclusion that Ruth and Talman were right: integrators are a more reliable simulation tool than matrix codes. On the other hand, the propagation of matrices and their subsequent analysis was a powerful way to integrate perturbation theory on the computer. Major design tools such as the code SYNCH or MAD all had the ability to propagate certain matrices. In the case of MAD, it could compute second-order matrices and certain quantities up to the third degree in the expansion variables.

Because of these conflicting requirements, most computer codes in accelerators, the CERN code MAD being the prime example, were really gigantic library managers. MAD would manage various sets of modules capable of computing all sorts of quantities without any particular emphasis on mathematical self-consistency. In analogy with economics, I dubbed this practice ‘horizontal integration or monopoly’. For example, these codes had formulae for ‘tune shifts with amplitude’, ‘chromaticities’, ‘radiation damping’ and a myriad of other objects which can be extracted by some normalization procedure on the Poincaré return map expressed as a Taylor series. At the same time, these codes could do simulations with complex elements thrown in and in the presence of imperfections: misalignments, mispowering, etc. Of course, the various modules, programmed by independent researchers under different assumptions, could not be expected to give fully reliable results under all extreme conditions simulated by the code itself.

With the advent of automatic differentiation, it became clear to me that it could be possible and actually easy to compute the necessary Taylor maps from our most reliable simulation model—the (symplectic) integrator. One would only need to rewrite perturbation theory for finite time (or  $s$ ) maps rather than the classical Hamiltonian approach used by accelerator physicists. Such a theory [40–42] would fit like a glove the computing and programming environment provided to us by modern languages. Modern programming languages such as C++ or FORTRAN90 allow us to define new types and extend ordinary

<sup>31</sup> Talman introduced a split dubbed ‘type IR’ in his code TEAPOT for the sake of powerful interaction quadrupoles. The emphasis was on getting some of the linear properties right with few steps.

operations and functions to these new types. In particular, we can create a Taylor type on which all the standard operations are allowed. Moreover we can also create a polymorphic type that can change from real (floating point) to Taylor series [44, 64, 65] at execution time.

These computer possibilities permit us to write an integrator, for example an explicit symplectic integrator, which operates on a polymorph using the syntax we would use on a regular floating point number. This integrator can then produce, thanks to this polymorphism, a Taylor map around any trajectory it is capable of computing.

Thus, very early on, I started proposing [43, 66, 67] the idea of a ‘vertically integrated code’. Of course, such a code could still have modules which cannot be self-consistent. The best example is a module computing certain collective effects resulting from the particle’s  $f$  interaction with themselves or their environment. While first-order results concerning such effects may depend on single particle dynamics (the integrator), it obviously sits outside the scope of single particle dynamics. So, by vertical integration, I meant that mathematical quantities computed self-consistently from single particle dynamics should be available to a user of a vertically integrated code.

Normal form theory on Taylor maps, symplectic integration applicable to the standard models inside MAD, and automatic differentiation for the easy production of Taylor maps were all available by 1987. I personally stated at that time that codes should be ‘vertically integrated’ around a symplectic integrator as their self-consistent central core [43].

Of course, as I pointed out, even to this day, the fact that this program requires ‘Taylor maps’ kept alive the fiction that I was still a strong proponent of Taylor map tracking. Nothing can be further from the truth. Obviously, schools of thoughts which still emphasize the Taylor map codes for tracking simulations happened to emphasize theoretical tools based on maps as well.

So, in summary, myself and others (Schmidt at CERN for example) have pushed for the application of symplectic integration to the basic magnets used in accelerators. We have claimed that *on the design trajectory*, the integrators would reproduce the Taylor matrices of the standard Taylor codes, and above all, of the biblical MAD code. Thus, around the year 1999, I started to develop the tools to overload the automatic differentiation of Berz in FORTRAN90. This library, called FPP<sup>32</sup>, contains the polymorphic types and the various normal form algorithms necessary to analyse Poincaré return maps. These analysis tools are the software implementation of the theory which generalizes the Courant–Snyder theory mentioned earlier—a Floquet theory of accelerator dynamics.

I subsequently wrote a library of magnet routines, dubbed PTC, which relies on FPP. PTC is notable for its implementation of symplectic integration and for a novel structure for the description of the beam line. In the rest of this section, I will summarize some of PTC’s tracking algorithms in the light of splitting methods.

### 5.1. TEAPOT split reviewed

The Hamiltonian of TEAPOT describes a magnet with cylindrical symmetry. The integration variable  $s$  is the arc of circle at the design radius  $\rho$ . We assume that the magnet is invariant along that arc. With this assumption, one can derive an expression for the most general potential and it will depend on  $\rho^{-1}$ . The general form for the Hamiltonian of this magnet is

<sup>32</sup> FPP, the fully polymorphic package is being documented on the web by E Forest and Y Nogiwa. It is completely independent of any tracking code. It can be used in any program: it simply facilitates the extraction of Taylor series and the normalization of Taylor maps. It is based on a version of Berz’s Package.

given by

$$\begin{aligned}
 H &= - \underbrace{\left(1 + \frac{x}{\rho}\right) \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}}_{T_1} + \underbrace{V(x, y; \rho^{-1})}_{T_2} \leftarrow \text{the TEAPOT split} \\
 &= - \underbrace{\left(1 + \frac{x}{\rho}\right) \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2} + b_1 \left(x + \frac{x^2}{2\rho}\right)}_{H_1} + \underbrace{V(x, y; \rho^{-1}) - b_1 \left(x + \frac{x^2}{2\rho}\right)}_{H_2}.
 \end{aligned} \tag{125}$$

As indicated in equation (125), there are at least two splits possible. First we have the drift-kick split represented by  $T_1$  and  $T_2$ . The Hamiltonian  $T_1$  represent a drift in polar coordinates and thus one computes the resulting map by pure geometry as Talman did in TEAPOT.

The second split is interesting and very useful. One notes that  $H_1$  is the Hamiltonian of a pure sector bend with a constant vertical field. If the strength  $b_1$  of the field is actually set to  $1/\rho$  then it would represent the ideal bend found in matrix codes. However, it turns out that  $H_1$  is exactly solvable for arbitrary  $b_1$ . The result is

$$\begin{aligned}
 x(s) &= \frac{\rho}{b_1} \left( \frac{1}{\rho} \sqrt{(1 + p_t)^2 - p_x(s)^2 - p_y^2} - \frac{dp_x(s)}{ds} - b_1 \right), \\
 p_x(s) &= p_x \cos\left(\frac{s}{\rho}\right) + \left( \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2} - b_1(\rho + x) \right) \sin\left(\frac{s}{\rho}\right), \\
 y(s) &= y + \frac{p_y s}{b_1 \rho} + \frac{p_y}{b_1} \left( \sin^{-1} \left( \frac{p_x}{\sqrt{(1 + p_t)^2 - p_y^2}} \right) - \sin^{-1} \left( \frac{p_x(s)}{\sqrt{(1 + p_t)^2 - p_y^2}} \right) \right), \\
 p_y(s) &= p_y, \quad p_t(s) = p_t, \\
 t(s) &= t + \frac{(1 + p_t)s}{b_1 \rho} + \frac{(1 + p_t)}{b_1} \left( \sin^{-1} \left( \frac{p_x}{\sqrt{(1 + p_t)^2 - p_y^2}} \right) - \sin^{-1} \left( \frac{p_x(s)}{\sqrt{(1 + p_t)^2 - p_y^2}} \right) \right).
 \end{aligned} \tag{126}$$

Incidentally one can extract from equation (126) the expression for the drift in polar coordinates by taking the limit  $\rho \rightarrow 0$ ,  $s \rightarrow 0$  and  $s/\rho = \theta$ :

$$\begin{aligned}
 x^{\text{new}} &= \frac{x}{\cos(\theta) \left(1 - \frac{p_x \tan(\theta)}{p_z}\right)}, & p_x^{\text{new}} &= p_x \cos(\theta) + \sin(\theta) p_z, \\
 y^{\text{new}} &= y + \frac{p_y x \tan(\theta)}{p_z \left(1 - \frac{p_x \tan(\theta)}{p_z}\right)}, & p_y^{\text{new}} &= p_y, \\
 t^{\text{new}} &= t + \frac{(1 + p_t) x \tan(\theta)}{p_z \left(1 - \frac{p_x \tan(\theta)}{p_z}\right)},
 \end{aligned} \tag{127}$$

where

$$p_z = \sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}.$$

### 5.2. The Cartesian bend

Another limit of the sector bend is the Cartesian bend. The Hamiltonian can be broken in several ways. Two convenient ways are given as follows:

$$\begin{aligned}
 H &= \underbrace{-\sqrt{(1+p_t)^2 - p_x^2 - p_y^2}}_{T_1} + \underbrace{V(x, y)}_{T_2} \\
 &= \underbrace{-\sqrt{(1+p_t)^2 - p_x^2 - p_y^2} + b_1 x}_{H_1} + \underbrace{V(x, y) - b_1 x}_{H_2}, \\
 V(x, y) &= \operatorname{Re} \left( \sum_{n=1}^{\infty} \frac{(ia_n + b_n)}{n} (x + iy)^n \right).
 \end{aligned} \tag{128}$$

In equation (128), the first split was integrable using the original interpretation of Ruth. The second split requires the Lie re-interpretation. The difference with a sector bend resides in the assumed symmetries of this magnet. Although, its main purpose is to bend the beam, it is a magnet which is approximately translationally invariant and therefore requires Cartesian variables. This particular magnet is never well presented in typical accelerator codes because most bends are viewed as being approximately invariant along an arc length. Nevertheless there are bends of this type in the real world.

Once more the map of  $H_1$  is exactly solvable. It can be obtained by taking a limit of equation (126). We take the limit  $\rho \rightarrow \infty$  and  $s = L$ :

$$\begin{aligned}
 x(L) &= x + \frac{1}{b_1} \left( \sqrt{(1+p_t)^2 - p_x(L)^2 - p_y^2} - \sqrt{(1+p_t)^2 - p_x^2 - p_y^2} \right) \\
 p_x(L) &= p_x - b_1 L \\
 y(L) &= y + \frac{p_y}{b_1} \left( \sin^{-1} \left( \frac{p_x}{\sqrt{(1+p_t)^2 - p_y^2}} \right) - \sin^{-1} \left( \frac{p_x(L)}{\sqrt{(1+p_t)^2 - p_y^2}} \right) \right) \\
 p_y(L) &= p_y, \quad p_t(L) = p_t \\
 t(L) &= t + \frac{(1+p_t)}{b_1} \left( \sin^{-1} \left( \frac{p_x}{\sqrt{(1+p_t)^2 - p_y^2}} \right) - \sin^{-1} \left( \frac{p_x(L)}{\sqrt{(1+p_t)^2 - p_y^2}} \right) \right).
 \end{aligned} \tag{129}$$

We should point out that the integration in Cartesian coordinates in a bend requires patching at the end: the frame must be rotated by an angle equal to half the design bending angle. This rotation is none other than the drift in polar coordinates of equation (127).

### 5.3. Straight magnets

When the magnet is a straight multipole magnet, we provide three splits. The Hamiltonian of the body of the magnet is again just that of equation (128). A straight magnet is characterized by a straight beam pipe running through the magnet. Its central purpose is usually not bending, but it is rather a focusing agent of some sort. Therefore, the standard Ruth split  $T_1 + T_2$  is retained. The second split of equation (128) is rather useless since  $b_1$  is often zero<sup>33</sup> and the

<sup>33</sup> In a straight magnet the pipes at the entrance and the exit are parallel and lined up. Therefore  $b_1 \approx 0$  is a necessity.

expressions in equation (129) are ill behaved for  $b_1 = 0$ . Instead it is useful to separate the quadratic part of the Hamiltonian from its nonlinear part. Thus we have the following split:

$$H = \underbrace{\frac{p_x^2 + p_y^2}{2(1 + p_t)} - p_t + \frac{b_2}{2}(x^2 - y^2)}_{H_1} + \underbrace{V - \frac{b_2}{2}(x^2 - y^2)}_{H_2} + \underbrace{C_{\text{exact}}}_{H_3}. \quad (130)$$

In equation (130), a nonlinear correction term is added. It is given by

$$C_{\text{exact}} = -\sqrt{(1 + p_t)^2 - p_x^2 - p_y^2} + \frac{p_x^2 + p_y^2}{2(1 + p_t)} + p_t. \quad (131)$$

Actually,  $H_1 + H_2$  is the old code SIXTRACK of Schmidt and Ripken. The additional kinematic correction is completely nonlinear in the angles and is a small correction in large storage rings. It is worth pointing out that this split will provide automatically the correct focusing properties of standard kick and Taylor codes. It is also for this split that the biased methods of section 4.5 are designed.

This is also the split used in the KEK<sup>34</sup> code SAD. SAD is a sophisticated accelerator tool which encompasses design, simulation and even control of an accelerator. It is even equipped with a Mathematica-like interface. The authors of SAD are definitely converts to symplectic integration and I quote here verbatim from the SAD home page:

Those who can stand with[sic] non-symplectic maps are not welcome to SAD (and to KEK).

The original SADists, as they called themselves, are K Hirata, S Kamada, K Oide, N Yamamoto and K Yokoya of KEK, Tsukuba, Japan.

Finally, I will point out that Hirata, Moshhammer and Ruggiero are also responsible for a symplectic beam-beam kick [22, 23] in the six-dimensional phase space. This kick is implemented by Schmidt in his code.

Leaving SADism, we conclude with another split used in PTC:

$$H = \underbrace{\frac{p_x^2 + p_y^2}{2} + \frac{b_2}{2}(x^2 - y^2)}_{H_1} + \underbrace{\left\{ -p_t \frac{p_x^2 + p_y^2}{2(1 + p_t)} - p_t + C_{\text{exact}} \right\}}_{H_2} + \underbrace{V - \frac{b_2}{2}(x^2 - y^2)}_{H_3}. \quad (132)$$

This split isolates the linear part and lumps all  $p_t$  dependence in the correction term. It has certain speed advantages.

Finally, all these three splits can be applied to the expanded Hamiltonian. For the expanded Hamiltonian  $C_{\text{exact}}$  is dropped and  $H_1$  is modified as follows:

$$H_1 \rightarrow H_1 - \frac{x p_t}{\rho_d} + \frac{x^2}{2\rho_d^2}. \quad (133)$$

In the expanded framework, no real difference is made between sector bends and Cartesian bends. The difference appears in certain ad hoc corrections located in the fringe regions, i.e., at the extremities of the magnet.

#### 5.4. The electric elements and non-symplectic radiation

I will not say much about electric elements (also known as radio frequency cavities) except to say that they are essential to maintain ‘longitudinal’ or time focusing. Particles within a

<sup>34</sup> KEK is the Japanese acronym for the High Energy Accelerator Research Organization.

bunch which are late/early compared to the centre of the bunch are put on an orbit with a smaller/larger circumference by changing the value of the energy  $p_t$ . In the case of a pill-box cavity, the general form of the Hamiltonian for the body is

$$K_{\text{cav}} = \underbrace{-\sqrt{(1+p_t)^2 - p_x^2 - p_y^2}}_{T_1} + \underbrace{\sum_{n=-N,N} A_n(r) e^{in\omega t}}_{T_2}, \quad \text{where } r = \sqrt{x^2 + y^2}. \quad (134)$$

The functions  $A_n$  are related trivially to the Bessel function  $J_0$ . This Hamiltonian is obviously amenable to explicit symplectic integration as shown in equation (134). Besides equation (134), we can also have a travelling wave cavity: the  $A_z$  component is multiplied by a factor of  $e^{i(\omega t - kz)}$ . These cavities are used in linear accelerators (linacs) to accelerate a beam. I will not discuss them here because they are not easily amenable to explicit symplectic integration due to a transverse vector potential which depends on time.

RF-cavities are also necessary in electron rings to restore the energy of the beam lost to radiation. This is not a symplectic effect: the actual map is nonsymplectic. This puts in question the *raison d'être* of this paper: why do we care about symplectic integration if the correct map is not symplectic?

The answer to this resides in two paradoxical and seemingly contradictory statements:

1. Radiation is a very small effect and therefore it makes sense to ignore it during a first analysis of a design. One can get a good idea of the dynamic aperture and stability of an electron ring by only taking into account Hamiltonian effects.
2. Radiation effects, both deterministic and quantum effects, are the primary factors determining the ultimate beam size of a bunch in an electron ring.

It is important to understand that the effect of radiation, although small, qualitatively changes the dynamics. For example, if we neglect nonlinear effects, we know that a stable linear symplectic system will have eigenvalues on the unit circle. If the eigenvalues are distinct, then small changes of the system will result in the eigenvalues moving on the unit circle. Unless they collide or reach the real axis, the system is stable under small perturbations. This property of symplectic matrices probably explains our ability to construct stable accelerators. It is clear that a simulation code should try to incorporate these features automatically: it is the case of a symplectic integrator.

In an electron ring, the RF cavities perform two fundamental tasks. First they provide longitudinal stability in the  $t - p_t$  plane as they do in proton machines. Secondly, they pump energy into the system which corresponds on average to the energy lost due to radiation. The original closed orbit without radiation is slightly altered. For the new closed orbit (fixed point of the map), the energy regained at the RF cavities is exactly equal to the energy lost through radiation. The matrix around the closed orbit is not symplectic: the eigenvalues are slightly inside the unit circle. Therefore, the beam starts to collapse towards the central fixed point until the quantum (photonic) nature of radiation becomes important and provides random diffusion. Thus the final size of the beam in an electron machine is the result of a competition between classical (deterministic) radiation and quantum effects.

As I said these effects are small, but as I have just argued, qualitatively important. Therefore, if we include radiation in a code, for example to compute the new eigenvalues of the monodromy matrix, then we must make sure that these small nonsymplectic effects are not buried under spurious violations of symplecticity. This explains why symplectic integrators are important in electron machines: the physically small nonsymplectic radiation, added on top of a symplectic integrator, can be seen easily even if the integrator uses few steps!

How does this work? Consider a beam element (magnet) represented by a Hamiltonian  $K$  and approximated in the tracking code by a time-reversible quadratic approximation  $S_2(s)$ . As we have seen this is general enough to encompass Ruth's integrator and the whole sequence of Yoshida's high-order formulae.

The effect of radiation can be added as a force  $\vec{F}^{\text{rad}}(\vec{x}, \vec{p})$  which we will assume to be  $s$ -independent within a particular magnet. Then it is clear that a new quadratic approximation of the Lie map can be written as

$$S_2^{\text{rad}}(\Delta s) = \exp\left(\frac{\Delta s}{2} \vec{F}^c \cdot \vec{\nabla}\right) S_2(\Delta s) \exp\left(\frac{\Delta s}{2} \vec{F}^c \cdot \vec{\nabla}\right). \quad (135)$$

The operators are all assumed to be in canonical variables as reflected by the superscript 'c'. A first-order solution of the transfer map associated with the operator  $\vec{F}^c \cdot \vec{\nabla}$  can be derived easily. We will proceed now with a sketch of such a derivation (see [55] or [68] for more details).

Following Sands [69] the change in the relative energy deviation  $p_t$  (we are assuming an ultra-relativistic electron) is given by

$$\frac{dp_t}{dt} = K_c (1 + p_t)^2 \frac{\vec{B}_\perp \cdot \vec{B}_\perp}{(p_0/e)^2}, \quad K_c = 1.407\,893\,57 \times 10^{-5} E_0^3. \quad (136)$$

The variable  $t$ , the time of flight times the speed of light  $c$ , is therefore just the path length in the case of ultra-relativistic electrons. The reference energy of the electron  $E_0$  is measured in GeV. The field  $\vec{B}_\perp$  is the component of the local magnetic field perpendicular to the direction of propagation. The quantity  $\vec{B}_\perp \cdot \vec{B}_\perp$  can be easily computed from the value of the field given along the unit vectors of a cylindrical frame of reference. Needless to say that these quantities should be available to a well-written tracking code.

Since our variable of integration is a distance  $s$  along the magnet, we need to convert the derivative with respect to the time  $t$  into a derivative with respect to  $s$ . This is done using the underlying Hamiltonian of the magnet under consideration:

$$\begin{aligned} \frac{dp_t}{ds} &= \frac{dp_t}{dt} \frac{dt}{ds} \\ &= \frac{dp_t}{dt} [t, K] = - \frac{dp_t}{dt} \frac{\partial H}{\partial p_t}. \end{aligned} \quad (137)$$

During the radiation process, an ultra-relativistic particle will emit a photon in the forward direction only. Thus, the usual directions  $\frac{dx}{ds}$  and  $\frac{dy}{ds}$  are left unchanged while the transverse momenta actually change. Therefore, to the symplectic integrator of step size  $\Delta s$ , we must add the radiative terms:

$$p_t^f = p_t + K_c (1 + p_t)^2 \frac{\vec{B}_\perp \cdot \vec{B}_\perp}{(p_0/e)^2} \frac{\partial H}{\partial p_t} \Delta s, \quad (138)$$

$$p_x^f = (p_x - a_x) \frac{1 + p_t^f}{1 + p_t} + a_x, \quad (139)$$

$$p_y^f = (p_y - a_y) \frac{1 + p_t^f}{1 + p_t} + a_y. \quad (140)$$

The above set of equations constitutes a first-order solution of the transfer map associated with the radiation process. Although it can be part of a multi-step Yoshida integrator, terms proportional to it will not converge with the rate predicted by the theory.

This situation can be improved by using the non-canonical variables  $(\frac{dx}{ds}, \frac{dy}{ds})$  instead of the canonical variables  $(p_x, p_y)$ . In non-canonical variables, the radiative operator has a simple form because only the energy  $p_t$  is changed by the process:

$$\vec{F}^{n-c} \cdot \vec{\nabla} = K_c(1 + p_t)^2 \frac{\vec{B}_\perp \cdot \vec{B}_\perp}{(p_0/e)^2} \frac{\partial H}{\partial p_t} \partial_s. \quad (141)$$

Furthermore the quantity  $\frac{\partial H}{\partial p_t}$  which expresses the change in path length cannot depend on the energy when expressed in terms of the non-canonical variables  $(\frac{dx}{ds}, \frac{dy}{ds})$ . Thus equation (135) can be rewritten as

$$\begin{aligned} \mathcal{S}_2^{\text{rad}}(\Delta s) &= \exp\left(\frac{\Delta s}{2} \vec{F}^c \cdot \vec{\nabla}\right) \mathcal{S}_2(\Delta s) \exp\left(\frac{\Delta s}{2} \vec{F}^c \cdot \vec{\nabla}\right) \\ &= \mathcal{C} \exp\left(\frac{\Delta s}{2} \vec{F}^{n-c} \cdot \vec{\nabla}\right) \mathcal{C}^{-1} \mathcal{S}_2(\Delta s) \mathcal{C} \exp\left(\frac{\Delta s}{2} \vec{F}^{n-c} \cdot \vec{\nabla}\right) \mathcal{C}^{-1}, \end{aligned} \quad (142)$$

where  $\mathcal{C}$  is a change of variables from  $(p_x, p_y)$  to  $(\frac{dx}{ds}, \frac{dy}{ds})$ . The operator  $\exp(\Delta s \vec{F}^{n-c} \cdot \vec{\nabla})$  changes only the variable  $p_t$  according to the relation:

$$\begin{aligned} p_t^f(\Delta s) &= \exp(\Delta s \vec{F}^{n-c} \cdot \vec{\nabla}) p_t \\ &= \left( \frac{1}{1 + p_t} - K_c \frac{\vec{B}_\perp \cdot \vec{B}_\perp}{(p_0/e)^2} \frac{\partial H}{\partial p_t} \Delta s \right)^{-1} - 1. \end{aligned} \quad (143)$$

Because equation (143) is an exact solution of the radiative operator of equation (142), if used in one of the high-order integrators previously discussed, it will behave appropriately and preserve the expected rate of convergence of the integrator.

### 5.5. Fringe fields

At the centre of most magnets, the fields are nearly constant along the time-like variable used to integrate. In fact, this variable is chosen so as to simplify the description of the fields in the interior of the magnet. However, at the ends of the magnet, the field must fall to zero. We refer to the fields of these regions as the ‘fringe fields’. Maxwell’s equations force the appearance of new terms which mess up our beloved explicit symplectic schemes.

In particular, fringe fields always pose a special problem for the correct non-expanded Hamiltonian. This leaves us with a few choices:

1. We can often revert to the expanded Hamiltonian and use the split of equation (110).
2. We can use implicit methods as Berg does with equation (111).
3. We can integrate using normal integrators and produce a Taylor map which is then symplectified. This is done in the Taylor codes. I already described this vaguely in the section on restoration (section 3.7.1).
4. We can kiss good bye to symplectic integrators. This is not a completely insane solution for electron or muon machines [70].
5. We can produce Taylor maps for the fringe region. This is perhaps an efficient way particularly if the fringe fields must be modified during a fitting algorithm. See for example Berz and Hoffstätter’s work [71] on the scaling of transfer maps in the fringe field region.
6. We can use a first-order impulse which can be made symplectic.

The last item of this list is done in most integrators in accelerators. Fringe effects are normally small in accelerators and therefore negligible. Typically they involve more powers

of our normally small momenta than a typical mapping for the core region of the magnet. Fringe effects also tend to cancel for straight magnets: the entrance field cancels the exit field. For this reason, most traditional codes such as the old kick codes include only the so-called vertical focusing of bending magnets. Worse, they only use the linear part of this vertical focusing!

However, in the original code TRANSPORT of Karl Brown, a fringe effect for the bends was included in the first- and second-order Taylor series. The results were published in the famous SLAC-75 technical report where the second-order analytic formulae of TRANSPORT [13] were first published. These fringe effects are also in the code MAD. Recently, as part of the insertion of my integrator PTC in the code MAD-X, I derived the correct results for large [72] excursions as well. The map of this fringe effect, at the entrance of the magnet, can be expressed with the following generating function:

$$F = p_x x^f + p_y y^f + \Delta t^f - \frac{1}{2} \Phi(p_x, p_y, \Delta) y^{f^2}, \quad (144)$$

where

$$\Phi(p_x, p_y, \Delta) = \frac{b_0 x'}{1 + y'^2} - g b_0^2 K \left\{ \underbrace{\frac{(1 - \Delta)^2 - p_y^2}{p_z^3}}_{[x, x']} + \underbrace{\frac{p_x^2}{x'^2}}_{x'^2} \underbrace{\frac{(1 - \Delta)^2 - p_x^2}{p_z^3}}_{[y, y']} \right\},$$

and  $\Delta = -p_t = -\Delta p / p_0$ . (145)

The field at the centre of the magnet is  $b_0 = \frac{B}{(p_0/e)}$ . The second-order terms are proportional to the constant  $K$  given by

$$K = \int_{-\infty}^{+\infty} \frac{b(z)(b_0 - b(z))}{g b_0^2} dz. \quad (146)$$

The constant  $g$  is the vertical pole gap. It does not really enter in the full expression because only the product  $gK$  is present. The quantities  $x'$  and  $y'$  are the drift space expression:

$$(x', y') = \frac{(p_x, p_y)}{\sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}}. \quad (147)$$

The reader will note that the map produced by this generating function is exactly solvable. This work is part of some unpublished documentation [72] of PTC/MAD-X.

Additionally, the first-order fringe field effects in a straight multipole can be computed and included as a generating function. This is described in a paper with Milutinovic [73] and also in [63]. It is also used in the code SAD of KEK. The most famous expression is the first order fringe field of the quadrupole first derived by Lee–Whiting [74] in 1970.

$$x^f = x \pm \frac{b_2}{12(1 + p_t)} \{x^3 + 3y^2x\}, \quad (148)$$

$$p_x^f = p_x \pm \frac{b_2}{4(1 + p_t)} \{2xyp_y - x^2p_x - y^2p_x\}, \quad (149)$$

$$y^f = y \mp \frac{b_2}{12(1 + p_t)} \{y^3 + 3x^2y\}, \quad (150)$$

$$p_y^f = p_y \mp \frac{b_2}{4(1 + p_t)} \{2xyp_x - y^2p_y - x^2p_y\}, \quad (151)$$

$$p_t^f = p_t, \quad (152)$$

$$t^f = t - \frac{f_{\pm}}{(1 + p_t)}, \quad (153)$$

where

$$f_{\pm} = \pm \frac{b_2}{12(1 + p_t)} \{y^3 p_y - x^3 p_x + 3x^2 y p_y - 3y^2 x p_x\}. \quad (154)$$

The ‘ $\pm$ ’ refers to the entrance and exit of the magnet and actually  $f_{\pm}$  is a first-order Lie generator for this map. A generating function can easily reproduce this result to first order and it will be of the first degree in the momenta. The same is true for all the straight multipoles. See for example [63].

These leading-order results are useful as a ‘switch’ in a simulation code. They allow us to check if the fringe effects are qualitatively important or are dwarfed by the multipoles introduced by the designer—the chromatic sextupoles for example. Of course if the fringe effects turn out to be important, one must perhaps go back and perform the most serious calculations mentioned at the beginning of this section.

The idea of refitting the known linear properties due to Talman is important, otherwise one will necessarily exaggerate the qualitative differences introduced by fringe fields. The reader should look at [75] for a case where the authors seem to have neglected to refit linear properties. Since important fringe fields do not slow down tracking with Taylor maps but make integrators grind to a halt in large machines, I will leave it to the reader to guess what kind of codes the authors of [75] advocate! Nevertheless the techniques they advocate, if fringe effects are truly needed after a Talman-like refitting, are worthwhile to learn and incorporate [71] in integrators.

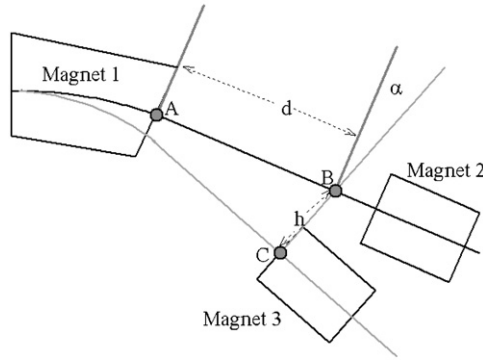
### 5.6. Euclidean group

In this section, we discuss issues of geometry which are not related to geometrical integration but are connected to the full exploitation of Darboux’s theorem which allows the creation of a map attached and belonging to a magnet.

The usage of Darboux theorem permits the passage from the normal relativistic Hamiltonian to maps describing the passage through a magnet. In the traditional beam optics point of view, a beam line is simply an ordered sequence of *different* magnets as we explained in section 2. The assumption is that the exit frame attached to a magnet is smoothly connected to the entrance frame of the subsequent magnet. If this were not the case, then equation (5) would be meaningless. In standard accelerator codes the geometry of the ring is deduced from the geometry of individual magnets. In fact most codes have the so-called survey command which locates the full beam line from the sequence described in the so-called lattice file. On the computer, the sequence is simply an array of magnet maps.

It should not take too much thinking to realize that this is mathematically restrictive. Let us look at figure 16. In this drawing which represents a piece of a recirculator, a particle with a high energy passes through magnet 1 and reaches magnet 2. Here we have assumed that the exit frame of magnet 1 is properly aligned with the entrance frame of magnet 2. After exiting from magnet 2, the beam makes a full turn and returns to magnet 1 with a smaller energy as it is decelerated through a linac before magnet 1. As a result, magnet 1, the so-called separator, bends the beam towards another beam line whose first magnet is denoted by no. 3.

If, as I pointed out, a proper computer representation of the beam line were to be a sequence of distinct objects, then the second time around, magnet 1 would have to be a ‘clone’ of the original magnet. This is already problematic because there is really only a single magnet at this location. Secondly, it is clear that exact cloning would not work since its exit frame would not be aligned with magnet 3 but with magnet 2.



**Figure 16.** Patching in a recirculator.

The solution to this problem is to make beam lines into an ordered sequence of ‘containers’ which I personally dubbed ‘fibres’. These containers label the discretized variable  $s$ . On the computer, these fibres are part of a linked list. Each fibre has a pointer to the actual magnet. In addition, each fibre contains elements of the dynamical representation of the Euclidean group whose purpose is to patch the frame of references between fibres.

In the example of figure 16, the second time around, magnet 1 sits in a different fibre, but it is the same magnet in the silicon brain of the computer. This new fibre points to the fibre containing magnet 3 rather than to the fibre containing magnet 2. The Euclidean patch, which was identity in the first case, contains now a nontrivial rotation of angle  $\alpha$  as well as a translation in both the longitudinal and transverse directions.

We list here the Lie operators of the dynamical Euclidean group and compared them with the usual Lie operators in the standard  $(x, p_x, y, p_y, z, p_z)$  phase space.

Time	Darboux	
$p_x \quad p_x$	$p_x$	
$p_y \quad p_y$	$p_y$	
$p_z \quad p_z$	$\sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}$	(155)
$L_x \quad yp_z - zp_y$	$y\sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}$	
$L_y \quad zp_x - xp_z$	$-x\sqrt{(1 + p_t)^2 - p_x^2 - p_y^2}$	
$L_z \quad xp_y - yp_x$	$xp_y - yp_x$	

In equation (155), we recognize three translations and three rotations. In the dynamical case (Darboux), the translation in the  $z$  direction is simply propagation in free space. The rotations around the  $x$  and  $y$  axes are drifts in polar coordinates: we recognize here the map used in the symplectic integration of a sector bend. Elements of the Euclidean group which interfere with the direction of propagation, namely the  $z$ -direction, are all drifts!

The reader can check that the Lie algebra of the original time-based operators and that used in the Darboux system are identical. In fact, the groups are locally isomorphic. The reader will note, looking at equation (127), that this isomorphism cannot be global as the  $x$  and  $y$  rotations diverge at angles of  $\pm\pi/2$ . This is expected since the Hamiltonian used in

describing magnets is not defined for trajectories going backwards in the magnet, i.e., having angles with respect to the integration axis beyond  $\pm\pi/2$ .

The linked list structures briefly described here as well as the representation of the Euclidean group embedded in the links (fibre) of this list insure that the map of each magnet behaves like an independent object, fully orientable in space. This is true whether the map is symplectic or not.

Some of the ideas presented here have inspired the new version<sup>35</sup> of ‘The Accelerator Description Exchange Format’ of R Talman and N Malitsky, more precisely what they called an ‘installed element’. This is perhaps the best attempt to put some order in the myriad of accelerator codes and models. It is distinguished by a disciplined structure based on the *element-algorithm-probe* framework. As I pointed out in connection with older programming efforts, computer structures were not derived from our understanding of the physics or mathematics but rather reflected our ignorance: managers of disconnected modules each encapsulating the wisdom of their creators. This is an exaggerated statement, but I do hope that these new structures will go beyond the previous attempts.

## 6. Conclusion and acknowledgments

Symplectic integration is now gaining ground in accelerator physics. My own work is being integrated into well-known codes such as MAD-X of CERN and Cornell’s BMAD. Other people, such as Talman and Malitsky, who support integration as the main reliable modelling tool, are working hard in providing some well-structured environments where physics provides the guiding principles. In the domain of FFAG, an implicit symplectic integrator is being used by Berg. Finally nonsymplectic integrators are also being used in electron and muon machines *with great care, vigilance and respect* towards the symplectic condition. And, on top of this, the tools of the Taylor map are being made available within the context of integrators: nothing is lost.

I have omitted the work of a lot of people through ignorance and also because of lack of space. For example I only alluded to the fact that one can average the Hamiltonian and then apply the original symplectic scheme of Ruth to this average Hamiltonian. This has been used in wigglers and undulators resulting in seemingly different schemes which are in reality one and the same.

I would like to thank Ron Ruth for long conversations over the phone. He helped me dig out the early history of his contribution. I also had some useful discussions with Dick Talman (Cornell), Ying Wu (Duke), Scott Berg, Johan Bengtsson (the creator of Taylor polymorphism), Nikolai Malitsky (Brookhaven), David Robin and Weishi Wan (LBNL) and finally David Douglas of Jefferson Laboratory. Here in Japan, I had fruitful interactions with Shinji Machida and Alexander Molodjontsev. It goes without saying that my recent implementations (FPP and PTC) were helped and supported by Frank Schmidt and Eric McIntosh of CERN.

Finally, I benefited a lot as did the entire community, from the efforts of Alex Dragt and Martin Berz. They have developed tools, theoretical and software, which permit us to compute and analyse Taylor maps. As I said, this reinforces the power of all integration tools.

## References

- [1] Laslett L J 1987 Concerning the  $y$ -growth exhibited by algebraic transformations *Technical Report Mura Report* 246, MURA, 1957 (It can also be found in LBNL PUB-616: selected works of L Jackson Laslett)

<sup>35</sup> ‘The Accelerator Description Exchange Format’ version 2.0 is part of the Unified Accelerator Libraries. There is a lot of documentation on the Internet including a set of lectures.

- [2] Laslett L J 1957 Round-off errors from fixed-point linear algebraic transformations computed by IBM-704 program 117 *Technical Report Mura Report 302*, MURA
- [3] Arnold V I 1978 *Mathematical Methods of Classical Mechanics* (New York: Springer)
- [4] 1985 *Nonlinear Dynamics Aspects of Particle Accelerators: Proc. Joint US-CERN School on Particle Accelerators (Sardinia)* (Berlin: Springer)
- [5] Ruth R D 1983 A canonical integration technique *IEEE Trans. Nucl. Sci.* **30** 2669
- [6] Courant E D and Snyder H S 1958 *Ann. Phys., NY* **3** 1
- [7] Weidemann H 1976 *Technical Report PEP Note 220*, SLAC
- [8] Weidemann H 1981 *Technical Report PEP Technical Memo 230*, SLAC
- [9] Wrulich A 1984 RACETRACK: a computer code for the simulation of particle motion *Technical Report DESY 84-026*
- [10] Schmidt F 1994 SIXTRACK, version 1.2: single particle tracking code treating transverse motion with synchrotron oscillations in a symplectic manner *Technical Report CERN SL/94-56 (AP)*, CERN
- [11] Ripken G 1984 *Technical Report 85-084*, DESY (This is the oldest reference on SIXTRACK)
- [12] Schachinger L and Talman R 1987 *Part. Accel.* **22** 35 (E Forest checked TEAPOT against the PSR lattice paper of Dragt for the appendix of this paper)
- [13] Brown K L 1982 A first and second order matrix theory for the design of beam transport systems and charged particle spectrometers *Technical Report SLAC Report 75*, SLAC
- [14] Berz M 2000 COSY INFINITY Version 8 *Technical Report* Michigan State University
- [15] Berz M, Hoffmann H C and Wollnik H 1987 *Nucl. Instrum. Methods A* **258** 402
- [16] Dragt A J and Finn J M 1976 Lie series and invariant functions for analytic symplectic maps *J. Math. Phys.* **17** 2215–27
- [17] Douglas D R, Forest E and Servranckx R V 1985 A method to render second order beam optics programs symplectic *IEEE Trans. Nucl. Sci.* **32** 2279
- [18] Erdelyi B and Berz M 2001 Towards accurate simulation of fringe field effects *Phys. Rev. Lett.* **87** 114302
- [19] Butcher J C 2003 *Numerical Methods for Ordinary Differential Equations* (Chichester, West Sussex: Wiley)
- [20] Berg J S, Warnock R L, Ruth R D and Forest E 1994 Construction of symplectic maps for nonlinear motion of particles in accelerators *Phys. Rev. E* **49** 722 (also see SLAC-PUB-6037 and 6164)
- [21] Warnock R L and Berg J S 1997 Fast symplectic mapping and long-term stability near broad resonances *Technical Report SLAC-PUB-7464*, Stanford Linear Accelerator Center  
Also appeared in 1996 *Proc. ICFA Workshop on Nonlinear and Collective Phenomena in Beam Physics (Archidosso, Italy, 2–6 Sept.) (AIP Conf. Proc.)*
- [22] Hirata K 1995 Analysis of beam–beam interactions with a large crossing angle *Phys. Rev. Lett.* **74** 2228–31
- [23] Hirata K, Moshhammer H and Ruggiero F 1992 A symplectic beam–beam interaction with energy change *KEK Preprint 92-117 A*  
Hirata K, Moshhammer H and Ruggiero F 1993 *Part. Accel.* **40** 205–28
- [24] Warnock R L and Ellison J A 1997 Convergence of a Fourier-spline representation for the full-turn map generator *Technical Report SLAC-PUB-7465*, Stanford Linear Accelerator Center  
Also appeared in 1996 *Proc. Conf. on Particle Beam Stability and Nonlinear Dynamics (Institute of Theoretical Physics, Santa Barbara, 3–5 Dec.) (AIP Conf. Proc.)*
- [25] Warnock R L and Ellison J A 1997 From symplectic integrator to Poincaré Map: spline expansion of a map generator in Cartesian coordinates *Technical Report DESY 97-163* and also SLAC-PUB-7465, Deutsches Elektronen-Synchrotron
- [26] Irwin J 1989 A multikick factorization algorithm for nonlinear maps *Technical Report SSC-228*, SSC Central Design Group
- [27] Abell D T 1995 Analytic properties and Cremona approximation of transfer maps for Hamiltonian systems *PhD Thesis* College Park, MD, USA
- [28] Earn D J D and Tremaine S 1992 Exact numerical studies of Hamiltonian maps iterating without roundoff error *Physica D* **56** 1
- [29] Vivaldi F 2006 The arithmetic of discretized rotations *Preprint*
- [30] Vivaldi F and Roberts J A 2003 Arithmetical method to detect integrability in maps *Phys. Rev. Lett.* **90** 034102-1
- [31] Abell D T, Schmidt F and McIntosh E 2003 Fast symplectic map tracking for the CERN Large Hadron Collider *Phys. Rev. (Spec. Top.—Accel. Beams)* **6** 064001-1
- [32] Gjaja I 1989 Monomial factorization of symplectic maps *Technical Report* Dynamical Systems and Accelerator Theory Group, Physics Department, University of Maryland
- [33] Rangarajan G 2003 Polynomial map factorization of symplectic maps *Int. J. Mod. Phys. C* **14** 847
- [34] Dragt A J 2004 *Lie Methods for Nonlinear Dynamics with Applications to Accelerator Physics* (Unpublished: available in draft form, MD, USA)

- [35] McLachlan R I and Quispel G R W 2004 Explicit geometric integration of polynomial vector fields *BIT Numer. Math.* **44** 515–38
- [36] Makino K 1998 Rigorous analysis of nonlinear motion in particle accelerators *PhD Thesis* Michigan State University
- [37] Smith L 1986 Effects of Wigglers and undulators on beam dynamics *ESG Tech Note-24*
- [38] Elleaume P 1992 *A New Approach to the Electron Beam Dynamics in Undulators and Wigglers: Proc. 1992 European Particle Accelerator Conference (Berlin)* Frontiers edn, p 661
- [39] Furman M 1985 Simple method to symplectify matrices *Technical Report* SSC-TM-4001, SSC Central Design Group
- [40] Dragt A J 1985 *Proc. 1984 Study on the Design and Utilization of the Superconducting Super Collider (Snowmass)* ed R Donaldson and J Morfin (New York: American Physical Society)
- [41] Forest E 1990 A Hamiltonian-free description of single particle dynamics for hopelessly complex periodic systems *J. Math. Phys.* **31** 1133 (originally, SSC-111, 1987)
- [42] Bazzanti A, Mazzanti P, Servizi G and Turchetti G 1988 *Nuovo Cimento* B 51
- [43] Forest E 1987 Canonical integrators as tracking codes (or How to integrate perturbation theory with tracking) *Technical Report* SSC-138, SSC-CDG (also part of lectures delivered at Fermi National Laboratory)
- [44] Forest E, Schmidt F and McIntosh E 2002 Introduction to the polymorphic tracking code *Technical Report* CERN-SL-2002-044, *KEK-Report* 2002-3
- [45] Peggs S G and Talman R M 1986 Nonlinear problems in accelerator physics *Ann. Rev. Nucl. Part. Sci.* **36** 287
- [46] Chao A W 1979 *J. Appl. Phys.* **50** 595
- [47] Chao A W 1981 *Nucl. Instrum. Methods* **180** 29
- [48] Forest E and Ruth R D 1990 Fourth-order symplectic integration *Physica D* **43** 105
- [49] Candy J and Rozmus W 1991 A symplectic integration algorithm for separable Hamiltonian functions *J. Comput. Phys.* **92** 230
- [50] Forest E 1992 Sixth-order Lie group integrator *J. Comput. Phys.* **99** 209
- [51] Yoshida H 1990 Construction of higher order symplectic integrators *Phys. Lett. A* **150** 262
- [52] Suzuki M 1992 General non symmetric higher-order decomposition of exponential operators and symplectic integrators *J. Phys. Soc. Japan* **61** 3015–9
- [53] Bandrauk A D and Shen H 1991 *Chem. Phys. Lett.* **176** 428
- [54] Iselin C *et al* 2005 The MAD8, MAD9 and MAD-X programmes (the most up-to-date version can be found at CERN and is available on the Internet) The symplectic integrator PTC of Forest, Schmidt and McIntosh is now embedded in MAD-X
- [55] Forest E, Reusch M F, Bruhwiler D and Amiry A 1994 The correct local description for tracking in rings *Part. Accel.* **45** 66
- [56] Yu Y K, Forest E and Robin D S Explicit symplectic integrator for  $s$ -dependent static magnetic field *Phys. Rev. E* **68** 2003
- [57] Sagan D 2005 The BMAD reference manual *Technical Report* revision: 5.5, Cornell
- [58] Forest E, Bengtsson J and Reusch M 1991 Application of the Yoshida–Ruth techniques to implicit integration and multi-map explicit integration *Phys. Lett. A* **5** 99
- [59] Berg J S 2005 Private communication
- [60] Neuffer D and Forest E 1989 A general formalism for quasi-local correction of multipole distortions in periodic transport systems *Phys. Lett. A* **135** 197
- [61] Ritson D 1990 SSCTRK: a particle tracking code for the SSC *Technical Report* SLAC-PUB-5302 and SSCL-305, Stanford Linear Accelerator Center (work performed at the now defunct SSC laboratory)
- [62] McLachlan R I 1995 Composition methods in the presence of small parameters *BIT* **35** 258–68
- [63] Forest E 1997 *Beam Dynamics: A New Attitude and Framework* (Amsterdam, The Netherlands: Harwood Academic)
- [64] Forest E and Schmidt F 2001 The ‘full polymorphic package’ (FPP) *ACM SIGPLAN Fortran Forum* **20** 12–7
- [65] Forest E and Schmidt F 2000 *Map Creation and Analysis via Overloaded Tools in Fortran 90: Proc. European Accelerator Physics Conference (Vienna, Austria)* p 1399
- [66] Forest E and Nishimura H 1989 *Vertically Integrated Simulation Tools for Self-Consistent Tracking and Analysis: Proc. 1989 Particle Accelerator Conference* p 1304 (also LBL-25961)
- [67] Forest E and Hirata K 1992 A contemporary guide to beam dynamics *Technical Report* 92-12, KEK
- [68] Ohmi K, Hirata K and Oide K 1994 From the beam-envelope matrix to synchrotron–radiation integrals *Phys. Rev. E* **49** 751
- [69] Sands M 1970 *Technical Report* SLAC-121, Stanford Linear Accelerator
- [70] Méot F 1999 The ray-tracing code Zgoubi *NIM A* **427** 353–6 (this code is a plain integrator for rings without special enforcement of symplecticity)

- [71] Hoffstätter G H and Berz M 1996 Symplectic scaling of transfer maps including fringe fields *Phys. Rev. E* **54** 109
- [72] Forest E and Leemann S 2002 Fringe effects in MAD: PART I. Second order fringe in MAD-X for the module PTC *Technical Report* KEK (unpublished work as of September 2005, available on request)
- [73] Forest E and Milutinovic J 1988 *Nucl. Instrum. Methods A* **269** 474
- [74] Lee-Whiting G E 1970 *Nucl. Instrum. Methods* **83** 232
- [75] Berz M, Erdelyi B and Makino K 2000 Fringe field effects in small rings of large acceptance *Phys. Rev. (Spec. Top.—Accel. Beams)* **3** 124001